UNIVERSITÉ DU LUXEMBOURG

PhD-FHSE-2025-018
The Faculty of Humanities, Education and Social Sciences

DISSERTATION

Defence held on 07/07/2025 in Esch-sur-Alzette
to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN *SCIENCES SOCIALES*

by

Ángela JIANG-WANG
Born on 11 November 1996 in Marbella (Malaga) (Spain)

THE SOCIAL INCENTIVES OF PROSOCIALITY

Dissertation defence committee

Dr Philippe VAN KERM, dissertation supervisor
Professor, Université du Luxembourg

Dr Sabrina TEYSSIER
Research Professor (Directrice de recherche), GAEL, University Grenoble Alpes

Dr Anja LEIST, Chairman
Professor, Université du Luxembourg

Dr Angela SUTAN
Professor, ESSEC Business School

Dr Francesco Pio FALLUCCHI
Senior Assistant Professor, University of Bergamo.

Acknowledgments

Before starting my PhD, in what now feels like a previous life, I conducted research in Consumer and Organizational Psychology. I was confident with experimental methods, but transitioning into the field of behavioral economics felt like learning an entirely new language. For a long time, I felt like a fish out of water, trying to catch up with theoretical models, economic conventions, concepts, and methods. Somehow, I made it to the other side, and now I have completed a thesis composed of economics papers.

For this feat, first and foremost, I would like to thank my supervisors, Daniele Nosenzo and Philippe Van Kerm, for their guidance, mentorship, and for teaching me how to conduct research in economics—something much easier said than done. Special thanks go to Daniele for his patience and for the many discussions during my first year, including debates he probably never would have encountered with an economics student—such as whether our experimental outcomes should be hypothetical or incentivized (coming from a more psychology-adjacent background, this was less obvious than it might seem). Special thanks to Philippe as well for always making the administrative procedures at UniLu seem so easy and effortless (though I'm sure they're not), and for giving me the opportunity to teach in his class, trusting me enough to let me design the content of a session and deliver a lecture on it.

One of my greatest sources of pride during this PhD is the feeling that I succeeded in acquiring a new language. For the same reason, I would like to thank my co-authors—Silvia Sonderegger, Bertrand Verheyden, and Francesco Fallucchi—for their guidance, patience, mentorship, and trust in me. I never imagined I would come to enjoy economic theoretical models as much as I do—especially after getting through that initial phase of intimidation. Finally, I would like to extend my sincere thanks to Manuel Muñoz, with whom I did not have the pleasure of co-authoring, but who nonetheless served as a mentor figure to me. The same goes for my former supervisor during my time as a research assistant, Elena Reutskaja. I am also grateful to my jury members—Anja Leist, Sabrina Teyssier, and Angela Sutan—for their time and feedback.

Another of my greatest sources of pride is that I was able to work on topics I'm truly passionate about. It makes me happy to know that during my PhD, I did the kind of research I genuinely want to pursue. Even more importantly, I feel that during this time I finally found my identity as a researcher and discovered the kind of projects I want to continue working on in the future. I somehow survived the job market while finishing my thesis, and I couldn't be happier about my next stop at the University of Pennsylvania and the projects I'll be working

on there with Cristina Bicchieri.

I have so many people to thank, in addition to those already mentioned, for accompanying me throughout my PhD journey and helping keep me afloat during my darkest hours. This includes—but is by no means limited to—support during the job market and throughout the (harrowing) process of completing my thesis ahead of the originally planned schedule. Let me try to do them justice.

My deepest thanks go to my 3E PhD cohort friends: Nathalie, Roxane, Stella, Reha, and Sakshi. To put it simply, I would not have survived this PhD without them. I mean it. I owe them my PhD. They were always there through the toughest and happiest moments, the first to give feedback or help pilot my studies whenever I needed a hand, and they patiently listened to countless hours of complaints and breakdowns. This is especially true for my fellow imposters, Nathalie, Rox, and Stella—their emotional and psychological support was endless. I feel incredibly lucky to have met you. Your friendship is one of the greatest gifts I take with me from this PhD. Thank you so much, guys.

And here I also want to include my friends Axelle and Victor. Thank you for making Luxembourg a place I could call home and for becoming my found family during this PhD. Thank you for being by my side throughout this crazy journey—making the darker times brighter and always sharing in my joy during the happier ones. As before, your friendship is one of the greatest treasures I take with me, and whenever I think of my time in Luxembourg, I will always think of you.

At the risk of sounding repetitive, as I had the privilege of having fantastic friends and a strong support network, I want to thank my friends Manu and Bruno. At certain points, my PhD really felt like a "sink or swim" situation, and you consistently helped me not to sink. Thank you for putting up with me, even from 1,000 km away.

As I start to run out of space, I'll quickly mention a few more names to whom thanks are due. If you're listed here, it means I am grateful for the helping hand you extended to me at some point during my PhD: Constance, Jonas, Thais, Giorgia, Marcelo, Jack, Blanka, Ellie, and Ellen.

I feel grateful and indebted to many people throughout my PhD. But to close this acknowledgements section, I would like to thank myself. Thank you for stepping out of your comfort zone, challenging your limits, staying curious, working hard, and being resilient. Thank you for never losing sight of why you do research in the first place. At the end of the day, remember: that's what truly matters.

Contents

Abstract

This thesis is divided in three chapters; each chapter addresses different research questions and work as standalone papers.

Chapter 1. Do monetary incentives damage social esteem? An empirical framework to detect crowding-out potential for prosocial behaviors.

Monetary incentives sometimes backfire and "crowd out" prosocial behavior. A standard explanation is that monetary incentives reduce the social esteem received from taking a prosocial action. However, because social esteem is not readily observable, there is a lack of direct empirical evidence confirming and measuring this mechanism. We propose a novel, portable, and incentive-compatible methodology to measure the social esteem received from prosocial behaviors. Our methodology is based on vignettes and incentivized second-order beliefs, and it provides a potential toolkit to identify when monetary incentives are more likely to backfire before implementing them. We run a high-powered pre-registered experiment with a UK sample (N = 5,368) within the context of COVID-19 vaccinations. Following our theoretical setup, we specifically focus on comparing (i) reward vs. penalty, (ii) small vs. large monetary incentives, and (iii) low vs. high baseline vaccination rates. We observe significant reduction in social esteem from monetary incentives, but only in the case of rewards. The size of the incentive and the baseline vaccination rate seem to matter little. Using our experimental results, we estimate how much an individual needs to value social esteem over money to be "crowded out" in our setup.

Chapter 2. You might be underestimating how sustainable others are.

Household consumption is responsible for nearly three-quarters of global carbon emissions, particularly in high-income countries, yet sustained behavioral change remains challenging. Since misperceptions about others' concern for climate change have been identified as key barriers to climate action, we study whether people also misperceive: (i) others' level of sustainable behavior, (ii) the level of sustainable behavior others deem appropriate, and (iii) others' support for policies that restrict or tax individual behavior. We address these questions in a well-powered, pre-registered online study with 1,292 Luxembourgers, using incentivized second-order belief elicitation. Participants reported their own behaviors, personal norms, and policy support, and estimated the same for others across the three most carbon-saving domains: consuming vegetarian meals, reducing home heating, and using public transportation. We find that participants systematically and substantially underestimate others' sustainability across all areas and behavioral domains. Replicating previous studies, participants also underestimate how much others donate to offset carbon emissions in an incentivized donation task. Consistent with pluralistic ignorance, underestimations of behaviors and policy support

are shared by both sustainable and unsustainable individuals, though not across all domains. Consistent with the false consensus effect, all underestimations are most pronounced among the least sustainable individuals, and underestimations of personal norms are exclusively driven by them. Finally, we show that social expectations strongly predict individuals' sustainable behaviors, personal norms, and support for restrictive policies, underscoring the potential of correcting misperceptions to promote behavioral change.

Chapter 3. Effectiveness of correcting misperceived social norms: A longitudinal experimental approach with sustainable behaviors and policy support

Misperceived social norms can impede progress toward sustainable behaviors and the policies that promote them. This study examines the potential of norm-based interventions to correct these misperceptions and foster lasting change in three key carbon-saving domains: vegetarian consumption, home heating, and public mobility. We conducted a three-wave longitudinal experiment over nine months with a well-powered Luxembourgish sample (N = 912). After identifying systematic underestimations of others' sustainable behaviors, personal norms, and policy support, we designed two targeted interventions: one providing information about actual levels of sustainable behavior and personal norms, and the other about actual levels of policy support within our sample. Our interventions successfully increased sustainable behaviors and policy support, while also updating participants' personal norms. Heterogeneity analyses based on prior beliefs and behaviors show that these effects were driven by positive belief updating and upward behavioral adjustment toward the new perceived norm, with no evidence of behavioral backfiring. This suggests that norm-based interventions may be most effective in situations of pluralistic ignorance or false consensus effects. The strongest treatment effects emerged in the domain of vegetarian consumption and persisted for at least three months. However, participants were less responsive to norm corrections in public mobility. After ruling out three potential explanations, we conclude that norm corrections may be less effective for behaviors that are costlier to adopt. Finally, we show that interventions promoting sustainable behaviors and policy support do not necessarily act as substitutes.

Co-author Statement

All chapters of this thesis have been co-authored. I took the lead on each chapter and am listed as the first author in all of them. Below is a detailed account of my contributions to each paper:[1]

Chapter 1. Do monetary incentives damage social esteem? An empirical framework to detect crowding-out potential for prosocial behaviors. Co-authored with Daniele Nosenzo (Aarhus University) and Silvia Sonderegger (University of Nottingham).

I developed the original research question, which was subsequently refined in collaboration with all co-authors. I conducted the literature review. Silvia Sonderegger developed the initial theoretical framework (not used in this chapter), which I then adapted into a simplified version that serves as the basis for this chapter. All authors contributed to the experimental design. I was responsible for programming the experiment using Qualtrics software and obtaining IRB approval and data protection approval. I also applied for and secured two internal grants to fund the study, totaling €7,000. Daniele Nosenzo and I co-authored the pre-registration. I managed the data collection, performed the data cleaning and data analyses, and produced the tables and figures. Finally, I wrote the entire chapter.

Chapter 2. You might be underestimating how sustainable others are and **Chapter 3. Effectiveness of correcting misperceived social norms: A longitudinal experimental approach with sustainable behaviors and policy support**. Co-authored with Francesco Fallucchi (University of Bergamo), Philippe Van Kerm (University of Luxembourg), and Bertrand Verheyden (Luxembourg Institute of Socio-Economic Research).

These chapters were part of the large-scale project *SOC2050* conducted in collaboration with "Luxembourg Strategie" at the Ministry of Economy in Luxembourg. Bertrand Verheyden drafted the *SOC2050* project proposal and collaboration agreement with the Ministry in collaboration with all co-authors and other researchers at LISER. In addition to the academic papers (two of which are presented in this dissertation) produced in the framework of this project, we prepared a policy report with recommendations for policymakers in Luxembourg, of which I am also a co-author. With regard to the two chapters presented here, I developed the original research questions and hypotheses (these were subsequently refined in collaboration with all co-authors) and I conducted the literature review. All authors contributed to the survey and experiment design. I programmed the implementation of the surveys (including the experimental setup) using Qualtrics software, together with another researcher at LISER and a research assistant.

Bertrand Verheyden and I co-authored the pre-registration and secured IRB approval and data protection approval. Bertrand Verheyden and I also worked with various staff at LISER

[1]All co-authors have reviewed and approved this co-authorship statement.

to translate the surveys into multiple languages, identify a suitable participant panel, and manage data collection and participant payments. Philippe Van Kerm, a research assistant, and I conducted the initial data cleaning and merging. Philippe Van Kerm also developed the weights used in the subsequent data analyses. I then performed the data analyses, produced the tables and figures, and wrote the entirety of both chapters.

Use of AI disclaimer

During the preparation of this thesis, I used GPT-3.5 from OpenAI (2023) for proofreading purposes and to enhance linguistic precision, style, and readability. I confirm that I independently authored the work and that the entire text of this thesis was written by me. After using the tool to check grammar and language quality, I carefully reviewed and edited the text as needed. I take full responsibility for the content and wording of the thesis.

Funding acknowledgment

I gratefully acknowledge financial support from the Luxembourg National Research Fund under the PRIDE program (PRIDE19/14233191-DTU-3E), from LISER CcEXPAR (Seed Grant 2023), and from the LISER Living Conditions Department (2023 Funding Call). Additionally, data collection for the studies in Chapters 2 and 3 was commissioned and funded by the Luxembourg Ministry of the Economy (SOC2050).

General Introduction

General Introduction

What motivates people to act in the interest of society, even when doing so involves a personal cost? Prosocial behaviors, such as engaging in pro-environmental actions or getting vaccinated against COVID-19, play a vital role in the functioning of societies. While much research has focused on intrinsic motivations and material incentives as determinants of prosocial behavior, this thesis investigates another important influence: social incentives.

Following Ashraf and Bandiera (2018), I define social incentives as those factors that influence behavior stemming from our interactions with others. As noted by Bicchieri (2016), individual preferences do not develop in isolation; rather, they are shaped by our expectations about what others do and what they approve of. In other words, our preferences are often conditional on the behavior and beliefs of others. This dissertation focuses on two closely related forms of social incentives: social esteem and social norms.

Social esteem, or social image, refers to how others evaluate our "type." When we perform an action that others can observe, we send a signal about the kind of person we are. As defined by Bursztyn and Jensen (2017), a "type" can refer to any attribute that matters to an individual, such as economic, social, or political characteristics. Ideally, we want others to see us as a socially desirable type rather than an undesirable one. In the context of prosocial actions, we want others to perceive us as prosocial (Bénabou and Tirole 2006). Social image concerns have proven to be powerful motivators of prosocial behavior, including blood donations (Lacetera and Macis 2010), voting (Funk 2010, DellaVigna *et al.* 2016, Ali and Lin 2013), and charitable giving (DellaVigna, List and Malmendier 2012, Andreoni and Bernheim 2009). Notably, Grimalda, Pondorfer and Tracer (2016) found that the desire to maintain a positive social image can serve as a stronger motivator for cooperation than the threat of punishment, highlighting the critical role of social image in sustaining prosocial behavior. This insight has even led researchers to investigate how to measure the welfare effects of social image (Butera *et al.* 2022).

Social norms, by contrast, are informal behavioral rules sustained by social expectations. Bicchieri (2005) even refers to them as "the grammar of society." According to Bicchieri (2005, 2016), a social norm exists when individuals prefer to conform to a behavioral rule, provided they believe that (i) most others conform to it (empirical expectations) and (ii) most

others believe they ought to conform (normative expectations). Empirical evidence shows that social norms are also important drivers of prosocial behaviors, including charitable giving (Bicchieri *et al.* 2022), proenvironmental behavior (Cialdini and Jacobson 2021, Saracevic and Schlegelmilch 2021), blood donations (Graf *et al.* 2023), and cooperative behavior (Dimant *et al.* 2024).

Although social esteem and social norms are distinct constructs, they remain closely intertwined. As noted by Gross and Vostroknutov (2022), a key reason people adhere to social norms is their concern for social image. Members of a community often seek to avoid ostracism or social rejection and prefer to engage in behaviors considered socially appropriate while avoiding those deemed inappropriate. However, social image does not serve as the sole motivator behind norm compliance. People frequently follow norms even when others cannot observe or sanction their behavior (Gross and Vostroknutov 2022). As we find in Chapter 3, social norms can shape individuals' personal norms, that is, what they personally believe to be the moral or appropriate code of conduct. This influence can lead them to comply with these norms even when their behavior is not observable, in order to avoid a negative self-image. Furthermore, social norms often shape what is considered socially desirable, which in turn influences social esteem (Deutchman *et al.* 2024, 2023). Still, social norms are not the sole determinant of social image; other factors that introduce uncertainty into an individual's motivations for engaging in prosocial actions can also alter it (Bénabou and Tirole 2006).

Hence, although social esteem and social norms are conceptually related and mutually influential, I have strong reasons to treat them as distinct types of social incentives. In this dissertation, I contribute to the literature by deepening our understanding of these two forms of social incentives and their role in shaping prosocial behavior. Across three chapters, I address the following central questions:

How can we measure social incentives?

The main contribution of Chapter 1 is the development of a novel methodology to measure the social esteem associated with prosocial behaviors, grounded in the theoretical framework of Bénabou and Tirole (2006). Empirical evidence has shown that social image concerns influence behavior, for example, by manipulating the observability of one's actions or the social rewards attached to them. However, to our knowledge, no reliable method exists

to quantify the social esteem attached to different prosocial behaviors. Developing such a measure represents a crucial first step for researchers interested in related questions, such as how social esteem varies across contexts, whether external factors can alter social esteem, and how it can compare in relative importance to other motivators, such as economic incentives. Chapter 1 further examines all of these questions. Although past researchers have introduced innovative methods to measure social norms (Krupka and Weber 2013, Bicchieri and Xiao 2009), these approaches do not allow us to specifically capture social image. Hence, Chapter 1 contributes to the literature by introducing a methodology designed specifically to measure social esteem.

In Chapter 2, we adapt the methodologies developed by Bicchieri *et al.* (2022) and Krupka and Weber (2013) to measure social expectations (perceptions about social norms) through surveys. This measurement provides comprehensive, quantifiable, and comparable data on the different normative beliefs related to a given prosocial behavior: empirical expectations (beliefs about what others do), normative expectations (beliefs about what others consider appropriate), and personal norms (beliefs about what one personally considers appropriate). This approach allows us to empirically examine, for example, whether empirical expectations associate more strongly with behavior than normative expectations, as explored in Chapter 2. It also enables us to detect whether individuals misperceive the prevailing social norms in their community, an insight central to the following research question.

Are people accurate in their beliefs about social incentives?

In Chapter 2, we study whether individuals hold accurate beliefs about pro-environmental social norms. This question is particularly relevant because inaccurate beliefs about others' thoughts and behaviors can lead people to conform to misperceived norms, which may result in undesirable behaviors (e.g. Bursztyn, González and Yanagizawa-Drott 2020, Andre *et al.* 2024). Two common biases often contribute to these inaccurate beliefs: pluralistic ignorance and the false consensus effect.

Pluralistic ignorance arises when individuals significantly misperceive the prevalence of

an opinion or behavior (Shamir and Shamir 1997, Prentice and Miller 1996). People may believe that most others support an opinion or behavior that, in reality, only a minority endorses. This misperception can lead individuals to act against their personal norms in order to align with what they mistakenly perceive as the majority view. For example, although most young married men in Saudi Arabia privately supported women working outside the home, they greatly underestimated the support of other similar men. These misperceptions significantly reduced their willingness to help their wives search for jobs (Bursztyn, González and Yanagizawa-Drott 2020).

The false consensus effect, in contrast, occurs when individuals overestimate how widespread their own opinions or behaviors are (Ross, Greene and House 1977). Those affected by this bias assume that others share their views or actions to a greater extent than they actually do (Mullen *et al.* 1985). For example, both climate change believers and deniers perceived their own opinion on climate change as the most common (Leviston, Walker and Morwinski 2013). Beliefs about others' views on climate change causally influence behaviors such as the willingness to discuss the issue with others (Geiger and Swim 2016) and support for climate policies (Andre *et al.* 2024).

In Chapter 2, we find that people underestimate how much others engage in proenvironmental behaviors, the level of engagement others consider appropriate, and the extent to which others support policies that restrict environmentally unsustainable behaviors. We also find that these underestimations are consistent with both pluralistic ignorance and the false consensus effect. Moreover, we find that these misperceived social expectations are strong predictors of individuals' own behaviors. These findings motivate the next research question.

How can we leverage social incentives in policymaking?

Chapter 3 builds on the findings from Chapter 2 and examines the potential of using social incentives as a tool for behavior change. We correct our study participants' misperceptions by informing them about the actual proenvironmental norms within their reference group. We find that our information interventions increase both proenvironmental behavior and policy support within our sample.

Although messages about norms are a popular policy intervention to promote prosocial

behavior (e.g. Schultz, Khazian and Zaleski 2008, Ferraro, Miranda and Price 2011, De Groot, Abrahamse and Jones 2013, Kormos, Gifford and Brown 2015, Sparkman *et al.* 2021, Hallsworth *et al.* 2017, Dur *et al.* 2021), they are not always effective (e.g. Griesoph *et al.* 2021, Richter, Thøgersen and Klöckner 2018, Gravert and Collentine 2021). As Bicchieri (2016) emphasizes, designing effective interventions such as norm-nudging requires a proper diagnosis of the causal influences on behavior within a given population. In Chapter 3, we find that norm-based interventions are most effective when targeted individuals initially underestimated the norm and subsequently updated their beliefs in a positive direction. These interventions also prove most successful when individuals previously behaved below the norm and adjusted their behavior upward. Conversely, the interventions are ineffective when participants already hold accurate information or overestimate the norm, or when their behavior already exceeds the norm. We additionally find that the effect of our information treatments can persist in the longer term, and that the elasticity of behavior in response to social incentives varies depending on the nature and cost of the behavior.

Another ongoing debate in the literature concerns whether behaviors and support for policies act as complements or substitutes. Some studies report that interventions promoting individual proenvironmental behaviors may reduce support for climate policies, possibly by shifting perceived relative importance from policies to personal actions (Werfel 2017). Other studies find no evidence of such negative spillover effects (Maki *et al.* 2019, Sparkman, Attari and Weber 2021). In Chapter 3, we contribute to this debate by showing that norm-based interventions promoting behavior do not undermine policy support. Similarly, we also show that norm-based interventions promoting policy support do not undermine behavior. Taken together, these findings offer valuable guidance to policymakers seeking to identify the conditions under which norm-based interventions are most likely to succeed, and how effective social incentives can be as a policy tool.

A second key question, which complements the investigation into using social incentives as a policy tool, concerns the external factors that can shift these incentives. Monetary incentives, widely employed by policymakers to promote prosocial behavior, sometimes backfire and fail to increase the desired behavior (e.g., Gneezy and Rustichini 2000a,b, Wollbrant, Knutsson and Martinsson 2022, Serra-Garcia and Szech 2023, Holmås *et al.* 2010, Mellström and Johannesson 2008, Gächter, Kaiser and Königstein 2025). According to the theoretical framework proposed

by Bénabou and Tirole (2006), this failure occurs because monetary incentives can crowd out social esteem.

In Chapter 1, we provide empirical evidence that monetary incentives can reduce the social esteem associated with prosocial actions. We also provide theoretical and empirical evidence that the likelihood of such crowding-out effects depends on the type of incentive used. Beyond contributing to the literature, we present our novel methodology as a practical tool for policymakers to assess the crowding-out potential of different monetary incentives before their implementation, by measuring the associated changes in social esteem. Our findings and our methodology can help scholars and policymakers design more effective interventions while minimizing indirect costs to social incentives.

Together, the chapters in this dissertation deepen our understanding of how social incentives shape prosocial behavior, which external factors can impact social incentives and how policymakers can harness social incentives to promote prosocial outcomes.

References

Ali, S Nageeb, and Charles Lin. 2013. "Why people vote: Ethical motives and social incentives." American economic Journal: microeconomics, 5(2): 73–98.

Andreoni, James, and B Douglas Bernheim. 2009. "Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects." Econometrica, 77(5): 1607–1636.

Andre, Peter, Teodora Boneva, Felix Chopra, and Armin Falk. 2024. "Misperceived social norms and willingness to act against climate change." Review of Economics and Statistics, 1–46.

Ashraf, Nava, and Oriana Bandiera. 2018. "Social incentives in organizations." Annual Review of Economics, 10(1): 439–463.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and prosocial behavior." American economic review, 96(5): 1652–1678.

Bicchieri, Cristina. 2005. The grammar of society: The nature and dynamics of social norms. Cambridge University Press.

Bicchieri, Cristina. 2016. Norms in the wild: How to diagnose, measure, and change social norms. Oxford University Press.

Bicchieri, Cristina, and Erte Xiao. 2009. "Do the right thing: but only if others do so." Journal of Behavioral Decision Making, 22(2): 191–208.

Bicchieri, Cristina, Eugen Dimant, Simon Gächter, and Daniele Nosenzo. 2022. "Social proximity and the erosion of norm compliance." Games and Economic Behavior, 132: 59–72.

Bursztyn, Leonardo, Alessandra L González, and David Yanagizawa-Drott. 2020. "Misperceived social norms: Women working outside the home in Saudi Arabia." American economic review, 110(10): 2997–3029.

Bursztyn, Leonardo, and Robert Jensen. 2017. "Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure." Annual Review of Economics, 9(1): 131–153.

Butera, Luigi, Robert Metcalfe, William Morrison, and Dmitry Taubinsky. 2022. "Measuring the welfare effects of shame and pride." American Economic Review, 112(1): 122–168.

Cialdini, Robert B, and Ryan P Jacobson. 2021. "Influences of social norms on climate change-related behaviors." Current Opinion in Behavioral Sciences, 42: 1–8.

De Groot, Judith IM, Wokje Abrahamse, and Kayleigh Jones. 2013. "Persuasive normative messages: The influence of injunctive and personal norms on using free plastic bags." Sustainability, 5(5): 1829–1844.

DellaVigna, Stefano, John A List, and Ulrike Malmendier. 2012. "Testing for altruism and social pressure in charitable giving." The quarterly journal of economics, 127(1): 1–56.

DellaVigna, Stefano, John A List, Ulrike Malmendier, and Gautam Rao. 2016. "Voting to tell others." The Review of Economic Studies, 84(1): 143–181.

Deutchman, Paul, Gordon Kraft-Todd, Liane Young, and Katherine McAuliffe. 2024. "People update their injunctive norm and moral beliefs after receiving descriptive norm information." Journal of Personality and Social Psychology.

Deutchman, Paul, Julia Marshall, Young-eun Lee, Felix Warneken, and Katherine McAuliffe. 2023. "Descriptive Norms Influence Children's Injunctive and Moral Norm Beliefs." Available at SSRN 4348267.

Dimant, Eugen, Michele Gelfand, Anna Hochleitner, and Silvia Sonderegger. 2024. "Strategic behavior with tight, loose, and polarized norms." Management Science.

Dur, Robert, Dimitry Fleming, Marten van Garderen, and Max van Lent. 2021. "A social norm nudge to save more: A field experiment at a retail bank." Journal of Public Economics, 200: 104443.

Ferraro, Paul J, Juan Jose Miranda, and Michael K Price. 2011. "The persistence of treatment effects with norm-based policy instruments: evidence from a randomized environmental policy experiment." American Economic Review, 101(3): 318–322.

Funk, Patricia. 2010. "Social incentives and voter turnout: evidence from the Swiss mail ballot system." Journal of the European economic association, 8(5): 1077–1103.

Gächter, Simon, Esther Kaiser, and Manfred Königstein. 2025. "Incentives crowd out voluntary cooperation: evidence from gift-exchange experiments." Experimental Economics, 1–32.

Geiger, Nathaniel, and Janet K Swim. 2016. "Climate of silence: Pluralistic ignorance as a barrier to climate change discussion." Journal of Environmental Psychology, 47: 79–90.

Gneezy, Uri, and Aldo Rustichini. 2000a. "A fine is a price." The journal of legal studies, 29(1): 1–17.

Gneezy, Uri, and Aldo Rustichini. 2000b. "Pay enough or don't pay at all." The Quarterly journal of economics, 115(3): 791–810.

Graf, Caroline, Bianca Suanet, Pamala Wiepking, and Eva-Maria Merz. 2023. "Social norms offer explanation for inconsistent effects of incentives on prosocial behavior." Journal of Economic Behavior & Organization, 211: 429–441.

Gravert, Christina, and Linus Olsson Collentine. 2021. "When nudges aren't enough: Norms, incentives and habit formation in public transport usage." Journal of Economic Behavior & Organization, 190: 1–14.

Griesoph, Amelie, Stefan Hoffmann, Christine Merk, Katrin Rehdanz, and Ulrich Schmidt. 2021. "Guess What. . . ?—How Guessed Norms Nudge Climate-Friendly Food Choices in Real-Life Settings." Sustainability, 13(15): 8669.

Grimalda, Gianluca, Andreas Pondorfer, and David P Tracer. 2016. "Social image concerns promote cooperation more than altruistic punishment." Nature communications, 7(1): 12288.

Gross, Jörg, and Alexander Vostroknutov. 2022. "Why do people follow social norms?" Current Opinion in Psychology, 44: 1–6.

Hallsworth, Michael, John A List, Robert D Metcalfe, and Ivo Vlaev. 2017. "The behavioralist as tax collector: Using natural field experiments to enhance tax compliance." Journal of public economics, 148: 14–31.

Holmås, Tor Helge, Egil Kjerstad, Hilde Lurås, and Odd Rune Straume. 2010. "Does monetary punishment crowd out pro-social motivation? A natural experiment on hospital length of stay." Journal of Economic Behavior & Organization, 75(2): 261–267.

Kormos, Christine, Robert Gifford, and Erinn Brown. 2015. "The influence of descriptive social norm information on sustainable transportation behavior: A field experiment." Environment and Behavior, 47(5): 479–501.

Krupka, Erin L, and Roberto A Weber. 2013. "Identifying social norms using coordination games: Why does dictator game sharing vary?" Journal of the European Economic Association, 11(3): 495–524.

Lacetera, Nicola, and Mario Macis. 2010. "Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme." Journal of Economic Behavior & Organization, 76(2): 225–237.

Leviston, Zoe, Iain Walker, and S Morwinski. 2013. "Your opinion on climate change might not be as common as you think." Nature Climate Change, 3(4): 334–337.

Maki, Alexander, Amanda R Carrico, Kaitlin T Raimi, Heather Barnes Truelove, Brandon Araujo, and Kam Leung Yeung. 2019. "Meta-analysis of pro-environmental behaviour spillover." Nature Sustainability, 2(4): 307–315.

Mellström, Carl, and Magnus Johannesson. 2008. "Crowding out in blood donation: was Titmuss right?" Journal of the European Economic Association, 6(4): 845–863.

Mullen, Brian, Jennifer L Atkins, Debbie S Champion, Cecelia Edwards, Dana Hardy, John E Story, and Mary Vanderklok. 1985. "The false consensus effect: A meta-analysis of 115 hypothesis tests." Journal of Experimental Social Psychology, 21(3): 262–283.

Prentice, Deborah A, and Dale T Miller. 1996. "Pluralistic ignorance and the perpetuation of social norms by unwitting actors." In Advances in experimental social psychology. Vol. 28, 161–209. Elsevier.

Richter, Isabel, John Thøgersen, and Christian A Klöckner. 2018. "A social norms intervention going wrong: Boomerang effects from descriptive norms information." Sustainability, 10(8): 2848.

Ross, Lee, David Greene, and Pamela House. 1977. "The "false consensus effect": An egocentric bias in social perception and attribution processes." Journal of experimental social psychology, 13(3): 279–301.

Saracevic, Selma, and Bodo B Schlegelmilch. 2021. "The impact of social norms on pro-environmental behavior: a systematic literature review of the role of culture and self-construal." Sustainability, 13(9): 5156.

Schultz, Wesley P, Azar M Khazian, and Adam C Zaleski. 2008. "Using normative social influence to promote conservation among hotel guests." Social influence, 3(1): 4–23.

Serra-Garcia, Marta, and Nora Szech. 2023. "Incentives and defaults can increase COVID-19 vaccine intentions and test demand." Management Science, 69(2): 1037–1049.

Shamir, Jacob, and Michal Shamir. 1997. "Pluralistic ignorance across issues and over time: Information cues and biases." Public Opinion Quarterly, 227–260.

Sparkman, Gregg, Bobbie NJ Macdonald, Krystal D Caldwell, Brian Kateman, and Gregory D Boese. 2021. "Cut back or give it up? The effectiveness of reduce and eliminate appeals and dynamic norm messaging to curb meat consumption." Journal of Environmental Psychology, 75: 101592.

Sparkman, Gregg, Shahzeen Z Attari, and Elke U Weber. 2021. "Moderating spillover: Focusing on personal sustainable behavior rarely hinders and can boost climate policy support." Energy Research & Social Science, 78: 102150.

Werfel, Seth H. 2017. "Household behaviour crowds out support for climate change policy when sufficient progress is perceived." Nature Climate Change, 7(7): 512–515.

Wollbrant, Conny E, Mikael Knutsson, and Peter Martinsson. 2022. "Extrinsic rewards and crowding-out of prosocial behaviour." Nature Human Behaviour, 6(6): 774–781.

Chapter 1

Do monetary incentives damage social esteem? An empirical framework to detect crowding-out potential for prosocial behaviors

Do monetary incentives damage social esteem? An empirical framework to detect crowding-out potential for prosocial behaviors

with Daniele Nosenzo (Aarhus University) and Silvia Sonderegger (University of Nottingham)

1.1 Introduction

Monetary incentives are a popular policy tool for promoting socially desirable behavior. However, ample global evidence suggests that they can have negative unintended effects. The most concerning one is the "crowding-out effect," which describes situations where incentives backfire and reduce people's willingness to engage in the desired behavior. In extreme cases, crowding out can lead to *fewer* people engaging in the behavior, sometimes with lasting effects even after incentives are removed (e.g. Gneezy and Rustichini 2000a,b, Wollbrant, Knutsson and Martinsson 2022, Serra-Garcia and Szech 2023, Holmås *et al.* 2010, Mellström and Johannesson 2008, Gächter, Kaiser and Königstein 2025). In less extreme cases, crowding out can still result in monetary incentives failing to promote the desired behavior, which can still impose significant costs on institutions (e.g., Panagopoulos 2013, Linardi and McConnell 2008, Dwenger *et al.* 2016).

A well-accepted theory explaining this phenomenon is that monetary incentives reduce the social esteem received from taking prosocial actions (Bénabou and Tirole 2006, Benabou and Tirole 2011). Following the literature, we define social esteem (hereafter also referred to as esteem) as "how prosocial others think you are".[1] People care about how they are perceived by others, and their behavior can often convey information about themselves. Behaving prosocially, consequently, can signal to others that they are a prosocial person. However, monetary incentives can spoil this signal by making people appear motivated by greed instead of motivated by prosocial values. This can reduce the esteem that people would otherwise receive from their actions. If incentives cannot fully compensate the loss of esteem, crowding out occurs.[2]

A key challenge in the literature is that esteem is not readily observable or measurable. Previous

[1] Many authors also use the term *social image*. The reason we chose *social esteem* is because it implies, specifically, a positive image value.

[2] Note that crowding out does not necessarily mean that the total change in participation within a given population will be negative. Crowding out occurs when some individuals who would have previously taken the prosocial action now choose not to. This effect depends on how much individuals personally value esteem and money. If a sufficient number of people are crowded out, the total change in participation can be either zero (monetary incentives fully substitute esteem) or negative. We will elaborate on this in more detail in Section 1.2.

behavioral research have provided indirect evidence of this mechanism by finding out that monetary incentives are less effective in settings with increased social image concerns, such as publicly observable settings.[3] However, as far as we know, there is a lack of empirical evidence that directly measures the damage of monetary incentives on social esteem. We contribute to the existing literature by proposing an incentive-compatible methodology that provides a quantifiable measure of changes in social esteem. This allows us to directly test the mechanism proposed by Bénabou and Tirole (2006). A novelty of our methodology is that it captures the total *net* esteem associated with a prosocial action. Taking a prosocial action has positive signaling value only if the person who takes it is perceived as more prosocial than someone who does not. Conversely, if taking the action does not lead to greater esteem than not taking it, there is no social advantage to doing so. This difference in esteem between taking and not taking the action is what we denote as the *net esteem* of an action, or, in other words, its signaling value. Crowding out only occurs when the net esteem is reduced.

Another advantage of our methodology is that it enables easy comparability. If monetary incentives have the potential to damage net esteem, a second important research question is whether this potential varies across different monetary incentives and environments. To answer this, we conducted an experiment using our methodology to test the effect of four different incentive schemes in two different environments. Concretely, we were interested in studying large and small incentives, penalties and rewards, and high and low baseline compliance. In fact, a major portion of the literature on incentives and image concerns only evaluate the effectiveness of small rewards. While larger incentives and penalties are understandably more difficult to implement, it is unclear whether they would impact net esteem the same way.

Theoretical evidence shows that large and small incentives can affect esteem differently (Bénabou and Tirole 2006, Benabou and Tirole 2011). Empirical evidence of their heterogeneous impact on net esteem can inform us when using smaller or larger incentives is most optimal. For instance, several empirical studies show that monetary incentives are less likely to backfire as their size increases, a notion commonly referred to as "pay enough or don't pay at all" (Gneezy *et al.* 2003, Gneezy and Rustichini 2000b, Wollbrant, Knutsson and Martinsson 2022, Serra-Garcia and Szech 2023). This strategy can be effective if large incentives do not reduce net esteem more than small incentives. Otherwise, its effectiveness may be difficult to evaluate a priori (in Section 1.2, we formally present

[3]For instance, Ariely, Bracha and Meier (2009) manipulated the observability of a volunteering task and found that monetary incentives increased effort only when behavior was private. However, Linardi and McConnell (2008) implemented a similar design and found no significant differences in the effectiveness of monetary incentives in both public and private settings, providing some mixed evidence. Exley (2018) found that volunteering levels were lower when monetary incentives were publicly observable, compared to when they were offered in private.

the condition under which using large incentives is not dominated).

More interestingly, Bénabou and Tirole do not distinguish between penalties and rewards in their theory and assume they impact esteem in the same way. However, in practice, we know that penalties and rewards can have very different behavioral effects, even when they are monetarily equivalent (Andreoni, Harbaugh and Vesterlund 2003, Hossain and List 2012, Homonoff 2018). These results are consistent with loss aversion: the notion that losses loom larger than equal-sized gains (Tversky and Kahneman 1979, 1991). Following the reasoning that loss aversion can result in rewards and penalties having different esteem-reducing potential, in Section 1.2 we extend Bénabou and Tirole's framework by incorporating monetary loss aversion to the utility of individuals. We then show how this can lead to rewards having a higher potential to reduce net esteem than penalties (and, by extension, to crowding out).

Finally, theory shows that environmental factors, such as the baseline compliance rate in the population, can influence the impact of monetary incentives on esteem (Bénabou and Tirole 2006). Additionally, empirical evidence shows that compliance rates in a group can affect its members' moral judgments and injunctive norm beliefs (Deutchman *et al.* 2023, Deutchman 2023). Hence, we aimed to test empirically whether monetary incentives interact with baseline compliance and whether this could be a relevant determinant of net esteem damage (and, thus, of crowding out).

Beyond our contributions to the literature on crowding out, monetary incentives, and social image concerns, our methodology also has direct practical applications for policymaking. It can serve as a predictive tool to evaluate the potential for crowding out before implementing a monetary incentive intervention. Practitioners often face significant uncertainty about the effectiveness of monetary incentives and which incentive scheme is optimal to adopt. In the worst-case scenario, monetary incentives can completely backfire, reducing public engagement in the behavior. Even when they do not, they may simply prove ineffective, creating a costly burden for institutions. Empirical studies examining crowding out and the conditions leading to it generally infer their conclusions ex-post (e.g., Gneezy *et al.* 2003, Mellström and Johannesson 2008, Wollbrant, Knutsson and Martinsson 2022). That is, they observe under which conditions monetary incentives typically backfire or fail to promote behavior and under which conditions they typically succeed. Usually, these conclusions are drawn after a monetary incentive has already been implemented and behavior has been impacted. However, it is essential for practitioners to find a way to resolve this uncertainty ex-ante, before implementing a monetary incentive and impacting a population's behavior (Bowles and Polania-Reyes 2012). Since our methodology is based on vignettes, it allows practitioners to measure changes in the net esteem of an action without influencing actual behavior. Hence,

our methodology can serve as a cost-effective toolkit for practitioners to predict the potential for crowding out and compare it across different incentive schemes. We show theoretically in Section 1.2 why a reduction in net esteem is a necessary condition for crowding out to occur. We also show in Section 1.5 an example of how policymakers can interpret and make use of the data collected through our methodology when evaluating whether to implement a monetary incentive intervention, and how to choose the most optimal one.

Although our methodology is adaptable to a wide range of prosocial behaviors, in our experiment, we chose to measure the impact of monetary incentives on the net esteem of receiving a COVID-19 vaccine. The discussion on whether monetary incentives can increase vaccination uptake has been popular both in academia and in practice. A large body of empirical studies examining the effect of incentives on COVID-19 vaccinations has been published since the outbreak of the pandemic, including in major outlets such as *Science*, *Nature*, and *PNAS* (e.g., Campos-Mercade *et al.* 2021b, Schneider *et al.* 2023, Klüver *et al.* 2021, Duch *et al.* 2023). A recent systematic review identified at least 38 studies published up until March 2022 (Khazanov *et al.* 2023), but even more have been conducted since then. Khazanov et al.'s findings show that not all studies find a positive effect of monetary incentives on uptake, and when they do, effect sizes tend to be small. Moreover, not a single study reviewed by Khazanov et al. used penalties. Despite this, policymakers mirrored scholars in their eagerness: governments all around the world have been implementing *both* monetary rewards and penalties of different size and duration to spur vaccination uptake (e.g., Muller 2021, BNS 2021, Parmar 2021, Paravantes 2022, AP 2021, The Local Italy 2022, Aranda 2021). Given the abundance of policy action plans and academic studies examining the effects of monetary incentives on vaccination behavior, we found it essential to delve into the potential for crowding out. Since prosociality motivates COVID-19 mitigating behaviors (Campos-Mercade *et al.* 2021a), we also found it highly likely that COVID-19 vaccinations would be influenced by social esteem concerns. If monetary incentives indeed damage the net esteem of taking a COVID-19 vaccine, this could result in undesirable lasting effects even after the incentives are removed (Gneezy and Rustichini 2000a, Gächter, Kaiser and Königstein 2025).

As described above, Section 1.2 presents our theoretical framework which is based on Bénabou and Tirole (2006)'s model with binary action choice. We incorporate monetary loss aversion in the utility of individuals, and predict that net esteem is more likely to be reduced with rewards (vs. penalties), smaller incentives (vs. large) and in environments with lower compliance rates (vs. high). We additionally show that net esteem damage is a necessary condition for crowding out. Furthermore, the direction of certain changes in esteem could directly confirm the presence of

crowding out. Net esteem loss arises from *motivation uncertainty*—the uncertainty about whether someone takes a prosocial action due to genuine prosocial values or greediness. This uncertainty significantly reduces the esteem received for complying.

The difference between penalties and rewards (and between large and small monetary incentives) in their potential for net esteem loss is explained through two channels: (1) *Motivation uncertainty* is expected to be greater for rewards (small incentives), and (2) in the presence of *motivation uncertainty*, abstaining becomes a more reliable signal of one's prosociality than complying. That is, in the presence of *motivation uncertainty*, we expect an abstainer to be judged more harshly with penalties (larger incentives) than with rewards, whereas we expect a smaller difference in judgment for a complier across incentive schemes. We also expect *motivation uncertainty* to be stronger in environments with lower baseline compliance.

Section 1.3 describes our pre-registered online experiment: a high-powered UK sample (N = 5,368) was presented with twenty different vignettes that described a man facing the choice to take a COVID-19 booster. Each participant only read one vignette version (n $\simeq$ 270). We manipulated the information presented to participants in three areas: i) the incentive scheme offered to the man for taking the vaccine (No incentive; Penalty of 100 GBP, Penalty of 15 GBP; Reward of 100 GBP; Reward of 15 GBP), ii) vaccination rates in the previous year (15% or 84%), and iii) whether the man ultimately decided to take the vaccine or not. We then elicited incentivized second-order beliefs about the likelihood that the man engaged in three different prosocial behaviors in the recent past, aiming to capture three different dimensions of prosociality: trustworthiness, honesty and altruism.

Section 1.4 presents our main results: we found that in the absence of monetary incentives, the net esteem of taking the COVID-19 booster was positive. We also found that when rewards were present, net esteem significantly decreased by an average of 40%. Penalties, on the other hand, did not significantly reduce net esteem. Additionally, the size of the monetary incentive did not matter: small incentives were as damaging as large ones. The baseline vaccination rate also made no difference. Our results are consistent with our model predictions: rewards have a higher potential of damaging net esteem and crowding out vaccination uptake than penalties. Additionally, our results also support "pay enough or don't pay at all": if the policymaker is considering implementing rewards, increasing the incentive size can make up for the decreased net esteem. On the other hand, we did not find support for "fine enough or don't fine at all": neither small nor large fines decreased net esteem significantly. Further analyses informed us that the damage towards net esteem was caused mainly by reductions in the esteem of taking the vaccine (which can also be understood as "honor", Bénabou and Tirole (2006), Butera *et al.* (2022)), whereas the esteem from abstaining (which

can also be understood as "shame") remained unchanged. This is consistent with our *motivation uncertainty* mechanism. Our data also confirmed our theory that *motivation uncertainty* is greater or equal for rewards than for penalties, and that abstaining becomes a more reliable signal in the presence of *motivation uncertainty*. Policy communications aimed at making "shame" a greater motivator for vaccination uptakes could help palliate the negative effects of rewards.

In Section 1.5 we discuss in more detail the implications of our results for the policymaker, and provide a simple estimation of how much an individual needs to value net esteem over money to be "crowded out" when a reward of 15 GBP is offered. We estimate that if agents are willing to pay more than 88 GBP to move from being considered neither likely nor unlikely to be prosocial by their peers to being considered very likely to be prosocial (or 176 GBP to move from being deemed very unlikely to be prosocial to very likely to be prosocial), then a reward of 15 GBP can induce defection. We also present a simple graph showcasing the "net incentive" individuals would actually receive when a reward of 15 GBP is offered, based on their relative valuation of net esteem compared to money.

Section 1.6 concludes.

1.2 Conceptual framework

For our theoretical framework, we consider a simplified version of Bénabou and Tirole (2006)'s model with binary action choice, where agents trade-off prosocial, monetary and social esteem concerns when deciding whether to engage in a prosocial action.

Consider an agent who must decide whether to perform ($a = 1$) or not perform ($a = 0$) a prosocial action, which is publicly observable. The action generates a social return normalized to 1 but also involves a cost c (capturing effort as well as other types of cost, for instance due to psychological factors) which is common to all agents in the same environment. For an agent i, the *net* utility from $a_i = 1$ over $a_i = 0$ is

$$\theta_i - c + S_1 - S_0 \tag{1.1}$$

The parameter $\theta_i \in (\theta_{\min}, \theta_{\max})$, with $\theta_{\min} \geq 0$, is a privately known parameter which characterizes i's prosociality. S_1 denotes the esteem the agent gets when performing the action, while S_0 denotes the esteem when not performing the action. We have

$$S_k \equiv E\left(\theta_i \mid a_i = k\right) \text{ for } k = 0, 1.$$

Following the literature, esteem depends on an observer's beliefs about the agent's type θ_i conditional on the agent's given action. $S_1 - S_0$ thus denotes the net esteem: the gain in esteem from choosing to take a prosocial action over not taking it. Note that, keeping everything else equal, (1.1) is increasing in θ_i. We can therefore identify a threshold $\widehat{\theta}$ such that all agents with $\theta_i > \widehat{\theta}$ select $a_i = 1$ and all agents with $\theta_i < \widehat{\theta}$ select $a_i = 0$. We can therefore express S_1 as $E(\theta_i \mid \theta_i > \widehat{\theta})$ and S_0 as $E(\theta_i \mid \theta_i < \widehat{\theta})$. In the absence of incentives, we generally expect that $S_1 - S_0 > 0$. This means that an agent taking the prosocial action is considered as more prosocial than an agent that does not take it. Throughout the text, we use the sub-/superscrips b, r and p to denote baseline, reward and penalty.

1.2.1 Introducing monetary incentives

Consider now a situation where a monetary incentive $t > 0$ is offered to all agents who select $a = 1$. The monetary incentive t can take the form of a reward ($j = r$) or a penalty ($j = p$). The net utility from $a_i = 1$ is

$$\theta_i - c + \eta_j \delta_i t + S_1 - S_0 \tag{1.2}$$

where $\eta_j \delta_i t$ is the utility agent i gets from the monetary incentive t. There are two key features to this model:

- All agents are loss averse in their monetary concerns: they allocate a higher value to avoiding a penalty than to receiving a monetarily equivalent reward. This is reflected in the loss aversion parameter η, where $\eta_r = 1$ and $\eta_p > 1$.
- Agents differ in their marginal utility of money δ. For tractability, we assume that there are only two groups: "greedy" (with share q) and "standard" (with share $1 - q$). Standard agents have $\delta = 1$ while greedy agents have $\delta > 1$. The distribution of θ in the two categories is the same.

We assume that greedy agents have a δ sufficiently large that they comply independently of their underlying prosociality. Intuitively, q reflects the share of individuals who are mainly motivated by "greed" when they take a given prosocial action.

The main intuition of the model is that monetary incentives dilute the signaling value of taking the prosocial action. Now there is *motivation uncertainty*: observers cannot distinguish between agents who take the action simply because they are greedy (in which case their expected θ is $E(\theta)$) and those who are motivated at least in part by their prosociality (in which case their expected θ is

$E(\theta_i \mid \theta_i > \widehat{\theta_j})$). Note that, keeping everything else equal, $\widehat{\theta}$ is decreasing in $\eta_j t$. Hence, we expect $\widehat{\theta_b} > \widehat{\theta_r} > \widehat{\theta_p}$.

S_1, in the presence of monetary incentives, becomes

$$S_1^j \equiv (1-\alpha_j) \underbrace{E(\theta_i \mid \theta_i > \widehat{\theta_j})}_{\text{esteem if } i \text{ is standard}} + \alpha_j \underbrace{E(\theta)}_{\text{esteem if } i \text{ is greedy}} \tag{1.3}$$

where $\alpha_j \equiv \frac{q}{q+(1-q)(1-F(\widehat{\theta_j}))}$ is the share of greedy agents among compliers. Note that S_0^j, on the other hand, remains as $E(\theta_i \mid \theta_i < \widehat{\theta_j})$. This is because there is no uncertainty whether an abstainer is greedy or standard.

Let's denote the loss in net esteem resulting from the *motivation uncertainty* (that is, the loss in net esteem provoked by the uncertainty of whether a complier is greedy or standard) as ϕ_j. We can also express ϕ_j as:

$$\phi_j = S^b - S^j$$

If $\phi_j > 0$, then net esteem has been damaged by the introduction of incentives.

We assume that, in the absence of *motivation uncertainty* (in a world with no greedy agents), net esteem would remain unchanged with monetary incentives because S_1^j and S_0^j would move towards the same direction following changes in $\widehat{\theta_j}$. In other words, the presence of *motivation uncertainty* is a necessary condition for incentives to reduce net esteem.

With this established, we can present three predictions:

Prediction 1a: *Rewards have a higher potential to reduce net esteem than penalties ($\phi_r \geq \phi_p$).*

This is due to $\widehat{\theta_r} > \widehat{\theta_p}$. Since agents are loss averse, penalties attract more standard compliers than rewards. There are two channels explaining why this can lead to $\phi_r \geq \phi_p$:

1. *Motivation uncertainty* is larger for rewards. This is because we expect the share of greedy among compliers to be higher for rewards than for penalties ($\alpha_r > \alpha_p$). From 1.3, we can see that S_1^j is decreasing in α_j. The larger the share of greedy among all compliers, the larger the *motivation uncertainty* and thus the larger the potential for net esteem loss.

2. Due to the presence of *motivation uncertainty*, as long as $\alpha_r \geq \alpha_p$ holds, we expect $(S_0^r - S_0^p) \geq (S_1^r - S_1^p)$.[4] Intuitively, this means that, in the presence of *motivation uncertainty*, we expect

[4]Note that even if α remains equal between rewards and penalties, rewards would still have a higher or equal net esteem loss than penalties. Because of loss aversion, we have $\widehat{\theta_r} > \widehat{\theta_p}$. Consequently, we have

abstaining to become a more reliable signal than complying. Someone who abstains in the presence of penalties will have a lower expected prosociality than someone who abstains in the presence of rewards. However, someone who complies in the presence of penalties may have a lower expected prosociality (in the case of being standard) or may have the same expected prosociality (in the case of being greedy).

The same rationale applies when comparing large and small monetary incentives within the same incentive type (reward or penalty).

Prediction 2a: *Small monetary incentives have a higher potential to reduce net esteem than large monetary incentives ($\phi_{small} \geq \phi_{large}$).*

Consider now different environments, with different compliance costs. Under our assumptions, a higher c corresponds to a higher $\widehat{\theta}_b$ and hence less baseline compliance. In environments with lower baseline compliance, we expect the share of greedy among compliers to be larger once monetary incentives are introduced. In other words, we assume that α is increasing in $\widehat{\theta}_b$.

Prediction 3a: *There is higher potential for net esteem loss when (c is high and hence) baseline compliance is low than when (c is low and hence) baseline compliance is high ($\phi_{low} \geq \phi_{high}$).*

1.2.2 Compliance and crowding out

Standard agents will be crowded out if:

$$\begin{aligned} &\textit{for a reward } t : t < \phi_r \\ &\textit{for a penalty } t : \eta t < \phi_p \end{aligned} \tag{1.4}$$

Definition: We say that an incentive j generates *image crowding out* when it decreases the compliance of standard agents: $\widehat{\theta}_j > \widehat{\theta}_b$.

The necessary condition for a monetary incentive to decrease total compliance is that it generates image crowding out. Image crowding out occurs when the loss in the net esteem associated with the prosocial action exceeds the monetary incentive. It is easy to see that $\phi > 0$, and thus $S^j < S^b$, is a necessary condition for image crowding out. We can also establish the following predictions from 1.4:

$S_0^r - S_0^p = \lambda$, with $\lambda \geq 0$. We established before that $E(\theta_i \mid \theta_i > \widehat{\theta_r}) - E(\theta_i \mid \theta_i > \widehat{\theta_p}) = \lambda$. Hence, following 1.3, and keeping α equal, $S_1^r - S_1^p = (1-\alpha)\lambda$. This solves $S^p - S^r = \alpha\lambda$, which is ≥ 0. Same rationale applies for large and small monetary incentives within the same incentive type.

Prediction 1b: *Rewards have a higher potential to generate image crowding out than penalties.*

Prediction 2b: *Small monetary incentives have a higher potential to generate image crowding out than large monetary incentives.*

Prediction 3b: *There is higher potential for image crowding out when (c is high and hence) baseline compliance is low than when (c is low and hence) baseline compliance is high.*

If there is image crowding out ($\widehat{\theta}_j > \widehat{\theta}_b$), the loss in net esteem can become even larger. This is because α becomes larger, thus potentially further decreasing S_1^j. Furthermore, since $\widehat{\theta}_j > \widehat{\theta}_b$, we would also expect $S_0^j \geq S_0^b$.

Finally, fixing the nature of the incentive (reward or penalty), we compare a small vs. a large incentive. If compliance with the small incentive is as high or higher than with the large incentive, the large incentive is clearly suboptimal. In this case, we say that it is *dominated*. The condition $\phi_{small} \geq \phi_{large}$ rules out that the large incentive is dominated.

1.2.3 Key takeaways

- Image crowding out is a necessary condition for a monetary incentive to lower total compliance.
- If net esteem with a monetary incentive is higher or equal than in baseline this rules out image crowding out.
- Penalties are less likely to generate net esteem loss than rewards. If we can exclude net esteem loss with rewards we can also exclude it with penalties, but not vice versa.
- Large incentives are less likely to generate net esteem loss than small incentives. If we can exclude net esteem loss with a small incentive we can also exclude it with a large incentive, but not vice versa.
- Net esteem loss is less likely under high baseline compliance than low baseline compliance. If we can exclude net esteem loss with low baseline compliance we can also exclude it with high compliance, but not vice versa.
- If a monetary incentive decreases net esteem, it generates an indirect cost that lowers its direct effect. The size of the net esteem loss quantifies this cost.
- Penalties generate higher compliance than rewards. The difference in net esteem with a penalty and a reward provides a measure of the penalty's compliance advantage.
- If net esteem with a large incentive is higher or equal than with a small incentive, this rules out that the large incentive may be dominated.

1.3 Experiment

We conceptualize social esteem (S_k) as the inferences others make about the agent's prosociality after observing the agent's actions. In order to measure this, we chose to elicit second-order beliefs—essentially, what people think others think. This approach is grounded in the idea that individuals decide whether to take an action based on their perception of what others will think of them. In situations where agents have perfect information about others' opinions, second-order beliefs should align perfectly with the average of all first-order beliefs. However, in a situation where agents have imperfect information about what others think (e.g., a situation of pluralistic ignorance; Sparkman, Geiger and Weber 2022, Bursztyn, González and Yanagizawa-Drott 2020, Andre *et al.* 2024), the actual motivator for people to act would be their second-order beliefs about esteem.

We also decided to elicit judgments about the prosociality of an hypothetical individual in a vignette. The obvious advantage of this approach is that it allows us to capture net esteem changes without actually having to face behavioral consequences. Conceptually, individuals' predictions about what others will think of them if they act in a certain way should perfectly match their predictions about what others will think about *any* person acting that way.

Our design also mitigates concerns about social desirability bias in participants' responses. Since participants are asked to guess others' responses rather than report their own—and are incentivized to do so—there is little risk that they will adjust their answers to appear more favorable.

The pre-registration of our experiment can be accessed here: `aspredicted.org/18V_9GD`. The study received ethical approval from the LISER Research Ethics Committee (project number 87).

1.3.1 Design and Procedure

We conducted an online vignette experiment where participants read about a hypothetical man who worked as a janitor in a senior care home in the UK and was eligible to receive the seasonal COVID-19 booster vaccine.[5][6] The vignette stated that the man one day received a letter from the local Council inviting him to receive the seasonal COVID-19 booster within a two-week period. The information subsequently displayed in the vignette varied according to the treatment condition to which the participants were allocated.

[5]We chose a sample based in the UK since vaccine hesitancy was low in the country at the time of the study. According to the latest survey conducted by UK Office for National Statistics (2021), based on adults in Great Britain, 96% reported positive sentiment towards a COVID-19 vaccine.

[6]At this point in time, only individuals at risk or in close contact with individuals at risk were elibible to receive the seasonal booster in the UK (NHS 2023).

The experiment followed a 5 (Incentive Scheme: No incentive, Large Penalty, Small Penalty, Large Reward, Small Reward) x 2 (Vaccination Rate: High, Low) x 2 (Action: Taking the vaccine, Not taking the vaccine) between-subject factorial design, resulting in twenty versions of the vignette in total. Each participant read only one version. Table 1.1 provides a summary of our conditions.

Table 1.1: Design

Treatments: Incentive schemes / Environments: vaccination rate	High vaccination rate (84%)		Low vaccination rate (15%)	
No incentive (baseline)	Takes vaccine	Does not take vaccine	Takes vaccine	Does not take vaccine
Large Penalty (100 GBP)	Takes vaccine	Does not take vaccine	Takes vaccine	Does not take vaccine
Small Penalty (15 GBP)	Takes vaccine	Does not take vaccine	Takes vaccine	Does not take vaccine
Large Reward (100 GBP)	Takes vaccine	Does not take vaccine	Takes vaccine	Does not take vaccine
Small Reward (15 GBP)	Takes vaccine	Does not take vaccine	Takes vaccine	Does not take vaccine

Our main treatments consisted of five different incentive schemes. In the *No incentive* (baseline) treatment, monetary incentives were not mentioned at any point in the vignette. In the other four treatments, the man in the vignette was told that he could receive a reward (penalty) of 100 (15) GBP, introduced by the Council this year, for taking the booster within the two-week deadline. He would receive a cheque (penalty notice) at home after his vaccination was (not) confirmed in the NHS registry.

Additionally, we manipulated the information displayed to participants in two more areas: 1) *Baseline vaccination rate*. The man in the vignette was informed that either 84% or 15% of eligible Council residents took the booster in the previous season. 2) *Action*. The man in the vignette either took or did not take the seasonal COVID-19 booster within the following two weeks.

We based our monetary incentive levels and vaccination rates on real-world information to maintain realism. Specifically, we used real COVID-19 vaccination rates for the third dose booster in the UK by May 2023 (UK Health Security Agency 2023).[7] We selected the vaccination rates from the local authority area with the highest rate (Milford & Lymington South), and the area with the lowest rate (Harehills South). The monetary incentive levels we implemented match actual amounts used in past government policies or field experiments.

Following an approach similar to Lane, Nosenzo and Sonderegger (2023) in their *Prosocial traits* study, we used the *opinion-matching method* (Bicchieri and Xiao 2009, Bursztyn, González and

[7]The original website we accessed does not exist anymore, but here is an example of how the page looked like: https://web.archive.org/web/20230114110603/https://coronavirus.data.gov.uk/details/interactive-map/cases

Yanagizawa-Drott 2020, Bicchieri *et al.* 2022) to elicit incentivized second-order beliefs about the prosociality of the man described in the vignette.

After reading one of the twenty possible vignette versions, a first batch of participants had to report their personal opinion (unincentivized) of the likelihood that the man in the vignette engaged in six different behaviors in the recent past, using an eight-point scale ranging from "Very unlikely" to "Very likely". Three of these behaviors captured prosocial domains: *trustworthiness* (keeping a promise made to a friend), *honesty* (returning the extra change to a cashier when the cashier accidentally gave them more change than they were due), and *altruism* (making a donation to a friend who ran the TCS London marathon for charity). The other three behaviors were filler questions to distract participants from the study's true objective, to mitigate experimenter demand effects. They were socially desirable behaviors, but unrelated to prosociality (exercising regularly to keep fit; keeping a healthy diet; and reading at least two books per month). The order of display of these six behaviors was randomized between participants.

Subsequently, a second batch of participants had to predict which was the most common answer given by the first batch to each of the six questions. This task was incentivized: participants were informed that we would pick randomly one out of every twenty respondents to be eligible to receive bonus payment. We would then randomly select one of the six questions, and eligible participants who correctly predicted the most common answer to that question would receive a 5 GBP bonus. Additionally, all participants answered a question designed to test their attention to the survey, as well as some basic demographic questions: gender, age, and annual personal income.

Regardless of the bonus payment, all participants in both studies received a show-up fee of 0.60 GBP for their participation in the experiment (which had a median duration of 4 minutes). Detailed experiment instructions can be found in Appendices A1.8 and A1.7.

1.3.2 Sample

In November and December 2023, we recruited 607 Prolific participants to elicit their first-order beliefs and 5,627 Prolific participants to elicit their second-order beliefs (after excluding participants who failed the attention question, as pre-registered). Before collecting the data for the second-order beliefs, we conducted a power analysis to determine the sample size we needed, and pre-registered it. We simulated a distribution of prosociality ratings using means and standard deviations from our first-order beliefs data (baseline condition), and tested the power to detect different treatment effect sizes. We used an approach similar to Campos-Mercade (2018). According to our power analysis,

with the current sample size we had 80% power to detect a reduction of net esteem of at least 30% (0.43 SD) when any kind of incentive is introduced, plus an additional minimum difference of 30% between incentive schemes (e.g., moving from Penalties to Rewards). All participants were located in the UK and indicated to be fluent in English. Since the data from the first-order beliefs batch is irrelevant for our analyses, subsequently we only describe the data from the second-order beliefs batch.

Since we pre-registered using gender, age and income as covariates in all our analyses, we dropped from the analysis 259 participants who indicated "I prefer not to say" to a question asking about their annual personal income.[8] Hence, the final sample size of our main study was N = 5,368, with roughly 270 participants allocated to each vignette version.

Appendix A1.1 provides details about the demographic distribution of our sample. We also provide the average demographic data of the UK population at the time of the study for comparison. Our sample is consistent with that of a representative sample in age, gender and income range.

1.4 Main results

We averaged the three second-order beliefs about trustworthiness, honesty and altruism to form a single index of prosociality, which is our empirical measurement of esteem (S_k). The answers, which were measured with an eight-point scale ranging from *Very unlikely* to *Very likely*, were recoded from -1 to 1 with equal spacing in between.[9] As shown in Table A1.3 in Appendix A1.2, the correlation of these three traits was positive and significant. We replicated all the analyses reported in this section for each of the three traits separately as a robustness check, which we include in Appendix A1.5. We found that the effects were strongest in beliefs about honesty, however all three traits had the same directional results.

We also show that the distribution of esteem ratings in our sample was not polarized in Appendix A1.3, which lowers concerns about our participants displaying vaccine-hesitant attitudes.

1.4.1 The model

Our main effect of interest is the potential reduction in net esteem caused by the presence of monetary incentives: ϕ_j. Net esteem represents the gain in esteem from taking a COVID vaccine

[8]We replicated the analyses including this set of dropped participants and our results remained identical.

[9]Since there was no middle point in the scale, we don't have "0" in our data, -0.25 is immediately followed by 0.25.

over not taking it: $S_1 - S_0$. We also wanted to study this effect in two different environments: high baseline vaccination rate and low baseline vaccination rate.

Hence, we pre-registered the following regression model for each vaccination rate level:

$$S_i = \alpha + \gamma TakesVaccine_i + \sum_{j=1}^{4} \beta_j T_i^j + \sum_{j=1}^{4} \delta_j (TakesVaccine_i \times T_i^j) + \sum_{n=1}^{3} \omega_n X_i^n + e_i \qquad (1.5)$$

S_i is participant i's evaluation of the expected prosociality of the man in the vignette, and $TakesVaccine$ is a dummy taking value 1 if the man in the vignette took the vaccine and 0 otherwise. T_i^j with $j \in [1, 4]$ are dummies for the four treatments with monetary incentives (No Incentive is the omitted base category), taking value 1 if the vignette contained the incentive scheme and 0 otherwise. X_i^n with $n \in [1, 3]$ are control variables representing gender, age and income; and e_i is the error term. γ measures the difference in expected prosociality between not taking and taking the vaccine in baseline, in other words, the net esteem of taking a vaccine without incentives: $S_1^b - S_0^b$. β_j measures the difference in expected prosociality between baseline and the respective monetary incentive when the man does not take the vaccine: $S_0^b - S_0^{incentive}$. Our main coefficients of interest are the four δ_j, which capture the change in net esteem after the respective monetary incentive is introduced: $(S_1^b - S_0^b) - (S_1^{incentive} - S_0^{incentive})$. In other words, δ_j captures ϕ_j.

1.4.2 Net esteem of taking a COVID-19 vaccine

As shown in Figure 1.1, net esteem was significantly above zero in all treatments. This means that, regardless of incentive scheme and vaccination rate, our sample deemed the man taking the vaccine as more prosocial than the man not taking it ($S_1 - S_0 > 0$).

This was also supported by our regression results, as shown in Table 1.2.[10] As pre-registered, we used the Benjamini-Hochberg method (Benjamini and Hochberg 1995) to adjust the p-values of our tests of interest and correct for Multiple Hypothesis Testing. As described in Section 1.4.1, γ informs us that receiving the vaccine in the baseline condition increased expected prosociality ratings by 0.45 points (1.09 SD) when the vaccination rate was HIGH, and by 0.41 points (0.99 SD) when the vaccination rate was LOW, both $p < 0.001$.

[10]We replicated all analyses excluding the control variables (gender, age, income) from the regression model and the results remained unchanged.

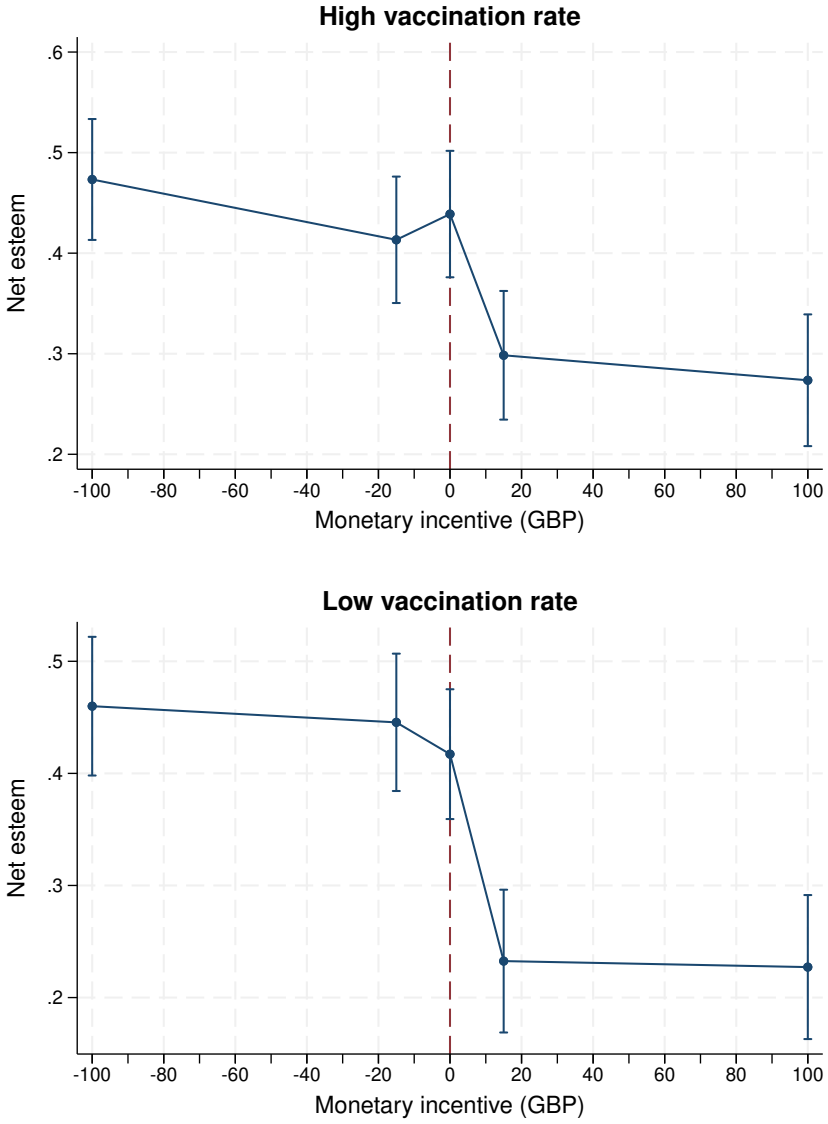


Figure 1.1: Net esteem ($S_1 - S_0$) by treatment condition. CIs at the 95% confidence level, estimated by calculating the standard error of the difference in means between S_1 and S_0 in each treatment condition.

1.4.3 Changes in net esteem across different incentive schemes

We can easily see in Figure 1.1 that rewards significantly decreased net esteem, regardless of size and vaccination rate. This was again confirmed by our formal analyses, as observed in the coefficients of the interaction terms $TakesVaccine \times LargeReward$ (δ_2) and $TakesVaccine \times SmallReward$ (δ_4) in Table 1.2.

The Large Reward treatment significantly decreased net esteem by 0.18 points (0.44 SD) for

Table 1.2: The effect of monetary incentives on expected prosociality, by vaccination rate

	High Vaccination Rate (1)	Low Vaccination Rate (2)
TakesVaccine=1 (γ)	**0.453*** **	**0.410*** **
Large Penalty (β_1)	-0.0477	-0.146***
Large Reward (β_2)	0.0558	-0.00476
Small Penalty (β_3)	-0.0707*	-0.122***
Small Reward (β_4)	0.0354	-0.0107
TakesVaccine=1 × Large Penalty (δ_1)	**0.0169**	**0.0507**
TakesVaccine=1 × Large Reward (δ_2)	**-0.183*** **	**-0.182*** **
TakesVaccine=1 × Small Penalty (δ_3)	**-0.0429**	**0.0439**
TakesVaccine=1 × Small Reward (δ_4)	**-0.165*** **	**-0.179*** **
Constant	-0.300***	-0.198***
Controls	Yes	Yes
Observations	2678	2690

* p<0.05, ** p<0.01, *** p<0.001. Adjusted p-values in bold.
Control variables: gender, age, income (not reported)

both vaccination rates ($p < 0.001$). The Small Reward treatment significantly decreased net esteem by 0.17 points (0.40 SD) when the vaccination rate was HIGH, and by 0.18 points (0.43 SD) when the vaccination rate was LOW (both $p < 0.001$). Hence, we can conclude that, for both incentive sizes, $\phi_r > 0$.

On the other hand, as shown in Figure 1.1, net esteem remained largely unchanged for penalties (compared to baseline). Accordingly, the coefficients of the penalty interaction terms δ_1 and δ_3 were non significant. Again, this was robust regardless of incentive size and vaccination rate. In other words, we could not reject that $\phi_p = 0$.

Wald tests to check whether $TakesVaccine \times LargePenalty = TakesVaccine \times LargeReward$ ($\delta_1 = \delta_2$) and whether $TakesVaccine \times SmallPenalty = TakesVaccine \times SmallReward$ ($\delta_3 = \delta_4$) confirmed that rewards were significantly more damaging to net esteem than penalties ($\phi^r > \phi^p$). We conducted these tests for both vaccination rates separately and all four tests had $p < 0.001$ except for $TakesVaccine \times SmallPenalty = TakesVaccine \times SmallReward$ with HIGH vaccination rate, which had $p = 0.01$.

Similar Wald tests comparing the effect of large and small incentives ($\delta_1 = \delta_3$ & $\delta_2 = \delta_4$) confirmed that the size of monetary incentives did not make any difference in our data, and that this was robust regardless of incentive type and vaccination rate.

Finally, we performed joint Wald tests to check whether $\gamma + \delta_1$, $\gamma + \delta_2$, $\gamma + \delta_3$ and $\gamma + \delta_4$ were equal to zero, which tells us whether net esteem remained significantly above zero in the presence of monetary incentives. Confirming what we already could see in Figure 1.1, we obtained $p < 0.001$ in all tests. Hence, albeit we found that rewards significantly damaged net esteem, the extent of this damage was not large enough to completely nullify the signaling value of taking the vaccine.

1.4.4 Testing our theoretical mechanisms

An independent analysis for S_1 and S_0 (details reported in Appendix A1.4) suggested that the damaging effect of rewards were mainly driven by a reduction of S_1, consistent with our hypothesized mechanism of *motivation uncertainty*. On the other hand, S_0 remained largely unchanged, which is consistent with the possibility of an *image crowding out* effect. On the other hand, both S_1 and S_0 were damaged (albeit not always significantly) for penalties, explaining the final net esteem remaining the same. Consistent with our theoretical mechanism of abstaining becoming a more reliable signal than complying, we confirmed that $(S_0^r - S_0^p) > (S_1^r - S_1^p)$. We also found that $S_1^p \geq S_1^r$, consistent with our hypothesis that *motivation uncertainty* is greater or equal for rewards than for penalties.

1.4.5 Discussion

In the absence of monetary incentives, we found a positive net esteem associated with taking a COVID-19 vaccine. Participants viewed someone who took the vaccine as more prosocial than someone who did not. However, we also found that introducing monetary incentives significantly diminished this net esteem. This detrimental effect was significant only for rewards, aligning with our theoretical predictions that rewards have a greater potential to reduce net esteem compared to penalties. Additionally, we found that small rewards were just as damaging as large rewards, and baseline vaccination rates had little impact. These findings are consistent with the patterns predicted by our theoretical framework.

We also found that the damaging effect of rewards was primarily driven by reductions in the "honor" associated with taking the vaccine (S_1), while the "shame" from not taking the vaccine (S_0) remained unaffected. These findings are consistent with the presence of *motivation uncertainty* and *image crowding out* as depicted in our theory. Additionally, we found that the difference in net esteem reduction between rewards and penalties could be explained by our two theoretical mechanisms: abstaining becoming a more reliable signal than complying, and *motivation uncertainty* being greater or equal for rewards than for penalties.

Finally, albeit rewards significantly damaged net esteem, the damage was not large enough to make it become zero. That is, even in the presence of incentives, taking a COVID-19 vaccine still had positive signaling value.

1.5 Implications for the policymaker

Our data reveals some interesting insights for UK policymakers, if they were to consider a monetary incentive policy to promote vaccinations. Consider the net return from compliance for a standard agent i in the baseline vs. the monetary incentives case. The net incentive from the monetary incentive, or the "effective t", is

$$\underbrace{\eta_j t}_{\text{direct incentive}} - \underbrace{\phi_j}_{\text{indirect net esteem effect}}$$

If $\phi_j > 0$, some of the money spent for incentivizing compliance is literally “eaten away”. Empirically, ϕ_j provides a measure of indirect loss in a given environment. Therefore, $\phi_r - \phi_p$ provides a measure of the indirect advantage of penalties over rewards.

A key implication of our findings is that, for a given t, taking the vaccine is less attractive with rewards than with penalties. This is driven by two channels: 1) the direct incentive of a reward is lower due to loss aversion ($\eta_p t > \eta_r t$), and 2) the indirect loss of a reward is higher ($\phi_r > \phi_p$).

A further implication of our data is that penalties seem to make vaccinations unambiguously more attractive for our sample: there is a direct incentive effect ($\eta_t t > 0$) with no unintended consequences on net esteem ($\phi^p = 0$). For rewards, on the other hand, the net incentive seems more ambiguous: whereas the money utility increases, the net esteem utility decreases. There is potential for crowding out.

The size of the incentive can be used strategically: If the policymaker is considering implementing rewards, the "pay enough or don’t pay at all" advice seems to be sensible here. Since large and small rewards were equally as damaging in our data, a larger reward can more easily compensate the decreased net esteem and reduce the risk of crowding out. On the other hand, we didn’t find support for "fine enough or don’t fine at all": since fines did not decrease net esteem, even a small fine could positively promote vaccinations.

A final suggestion to the policymaker is that a monetary incentive policy could be complemented with a communication policy. Since the damaging effect of rewards was mostly driven by penalizations to the "honor" received from taking the vaccine (consistent with *motivation uncertainty*), policy

communications aimed at increasing the social disapproval for skipping booster vaccines (hence making "shame" a greater motivator for vaccination uptake) could potentially attenuate the negative effects of rewards on net esteem.

1.5.1 Tipping point for image crowding out.

Let's consider an individual i that has $\theta_i = \hat{\theta}_b$ and they choose to take the vaccine without monetary incentives. Introducing a monetary reward of 15 GBP would reduce net esteem by 36% if the baseline vaccination rate were 84% and by 44% if the baseline vaccination rate were 15% (calculated from the coefficients in Table 1.2). For simplicity, let's assume an average reduction in net esteem of 40%. Given an average baseline net esteem of 0.43 (calculated as the average of the baseline net esteem for both vaccination rates), this implies that introducing a 15 GBP reward would reduce net esteem by 0.17 points.

Let's consider a parameter γ that represents the relative value standard agents place on net esteem compared to money. In other words, γ indicates how much standard agents value one unit of esteem gain in GBP. For individual i to choose to defect once a reward of 15 GBP is introduced, the following condition must hold:

$$\underbrace{15}_{\text{direct incentive}} - \underbrace{\gamma 0.17}_{\text{indirect net esteem effect}} < 0$$

This leads us to $\gamma > 88$. If a one-point increase in esteem is worth more than 88 GBP, then a reward of 15 GBP can induce defection. In more practical terms, a γ of 88 means, for instance, that individuals are willing to pay 88 GBP to move from being deemed neither likely nor unlikely to be prosocial to being considered very likely to be prosocial by their peers, or 176 GBP to move from being deemed very unlikely to be prosocial to very likely to be prosocial.

Moreover, we show in the y-axis of Figure 1.2 which would be the "effective t" in GBP for different levels of γ.

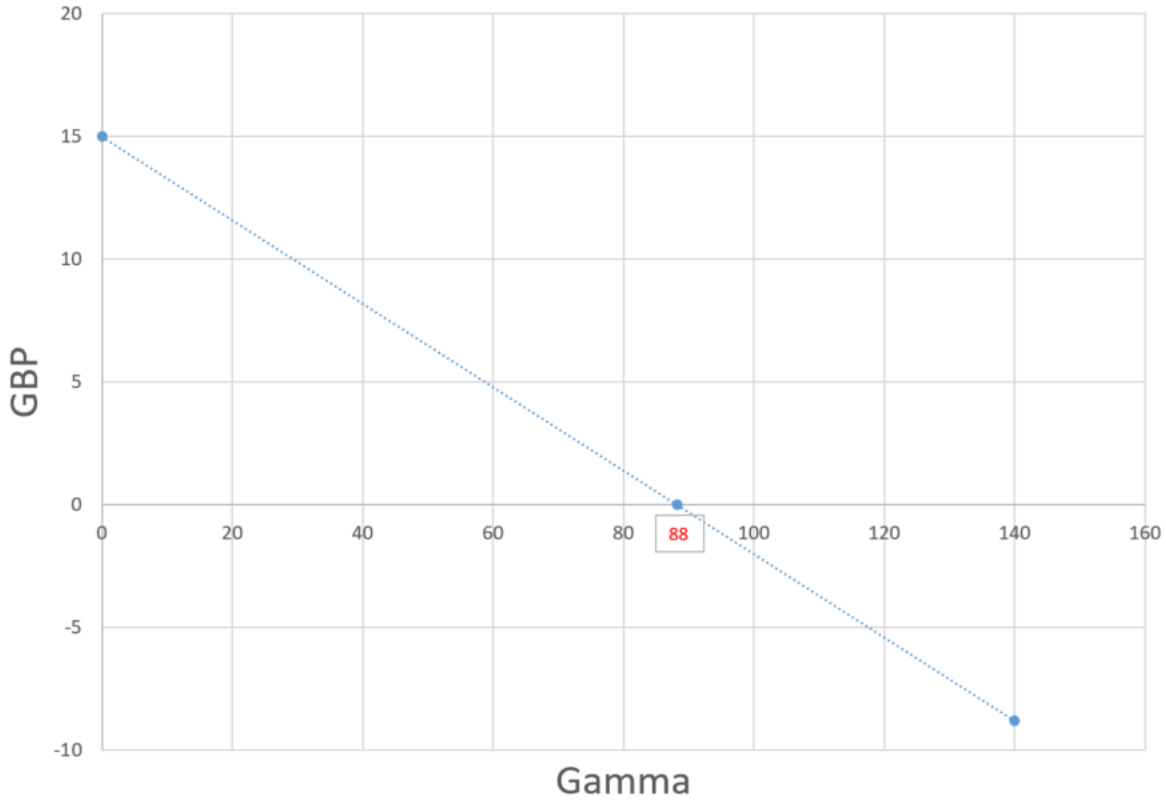


Figure 1.2: "Effective t" for a reward of 15 GBP

If $\gamma = 0$ (agents only care about money), the effective t they would be receiving with a 15 GBP reward would be the full amount. However, if $\gamma = 60$, the effective t agents would be receiving would actually be 5 GBP, since the loss of net esteem would be equivalent to losing 10 GBP to them. From $\gamma = 88$ onwards, the effective t becomes a loss.

1.6 Conclusion

We presented a novel and portable methodology to measure whether monetary incentives reduce the social esteem gained from doing a prosocial action. We conducted an online incentivized vignette experiment within the context of a COVID-19 booster and compared rewards against penalties; large against small incentives; and high against low baseline vaccination rates.

Our results showed direct empirical evidence of the mechanism proposed in Bénabou and Tirole (2006)'s model: that monetary incentives can reduce the net esteem associated with a prosocial action. However, we also show that, in our case, this was only true for rewards. Hence, we confirm that rewards and penalties can have different crowding-out potential and that their effects should be

evaluated separately by the policymaker. Taking our empirical results and our theoretical framework altogether, we recommend the use of penalties before that of rewards to reduce the risk for crowding out.

Whenever this is not possible, we recommend policymakers to implement larger rewards. In our results, small rewards were already enough to produce the damaging effect and, in fact, the size of the incentive seemed to matter little for net esteem reduction.

A standing problem of monetary incentives and crowding out is that it is very hard to predict ex-ante whether a monetary incentive will backfire. This can depend on many factors such as the nature of the prosocial behavior itself, the target population, and the type and size of the incentive. Our methodology has the advantage of allowing the measurement of net esteem damage without inducing changes in behavior. Hence, our methodology can be a possible cost-effective and easy-to-implement "workhorse" for policy-makers to evaluate crowding-out potential when considering implementing a monetary incentive policy.

The use of monetary incentives to promote prosocial behaviors, particularly in the context of COVID-19 vaccinations, has sparked debate among researchers, policymakers, and the general public. While previous studies have reported no adverse effects of offering monetary rewards for COVID-19 vaccinations on attitudes such as civic responsibility, self-determination, trust in providers, and perceived vaccine safety and efficacy (Schneider *et al.* 2023), our findings challenge this perspective by providing evidence that monetary incentives can negatively impact social esteem. Our results may also shed light on the widespread resistance to adopting monetary incentives as a tool for influencing health-related behaviors, both among policymakers and the general public (Campos-Mercade *et al.* 2024).

Future research could examine whether our results replicate for different behaviors and populations. Future studies could also examine different incentive levels -both smaller and larger than ours- to see whether the damaging effect of incentives remain unchanged. This could speak to whether the effect of incentives is marginal (depends on the size of the incentive), categorical (the effect is affected by the mere presence of the incentive) or a combination of both, as discussed by Bowles and Polania-Reyes (2012).

References

Andreoni, James, William Harbaugh, and Lise Vesterlund. 2003. "The carrot or the stick: Rewards, punishments, and cooperation." American Economic Review, 93(3): 893–902.

Andre, Peter, Teodora Boneva, Felix Chopra, and Armin Falk. 2024. "Globally representative evidence on the actual and perceived support for climate action." Nature Climate Change, 1–7.

AP. 2021. "Austria anti-vaxxers will be hit with €3,600 fine for refusing COVID-19 jab." Euronews. https://web.archive.org/web/20240109013519/https://www.euronews.com/2021/12/10/austria-anti-vaxxers-will-be-hit-with-3-600-fine-for-refusing-covid-19-jab.

Aranda, Rafa. 2021. "Galicia podrá sancionar con multas de hasta 60.000 euros a quien se niegue a vacunarse." Diario AS. https://web.archive.org/web/20231128040458/https://as.com/diarioas/2021/02/23/actualidad/1614103789_898986.html.

Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. "Doing good or doing well? Image motivation and monetary incentives in behaving prosocially." American economic review, 99(1): 544–555.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and prosocial behavior." American economic review, 96(5): 1652–1678.

Benabou, Roland, and Jean Tirole. 2011. "Laws and norms." National Bureau of Economic Research.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." Journal of the Royal statistical society: series B (Methodological), 57(1): 289–300.

Bicchieri, Cristina, and Erte Xiao. 2009. "Do the right thing: but only if others do so." Journal of Behavioral Decision Making, 22(2): 191–208.

Bicchieri, Cristina, Eugen Dimant, Simon Gächter, and Daniele Nosenzo. 2022. "Social proximity and the erosion of norm compliance." Games and Economic Behavior, 132: 59–72.

BNS. 2021. "Lithuanian government approves 100-euro payments for seniors who get vaccinated." Lithuanian Radio and Television (LRT). https://web.archive.org/web/20240531022145/https://www.lrt.lt/en/news-in-english/19/1514375/lithuanian-government-approves-100-euro-payments-for-seniors-who-get-vaccinated.

Bowles, Samuel, and Sandra Polania-Reyes. 2012. "Economic incentives and social preferences: substitutes or complements?" Journal of economic literature, 50(2): 368–425.

Bursztyn, Leonardo, Alessandra L González, and David Yanagizawa-Drott. 2020. "Misperceived social norms: Women working outside the home in Saudi Arabia." American economic review, 110(10): 2997–3029.

Butera, Luigi, Robert Metcalfe, William Morrison, and Dmitry Taubinsky. 2022. "Measuring the welfare effects of shame and pride." American Economic Review, 112(1): 122–168.

Campos-Mercade, P. 2018. "Power analysis through simulations in Stata: a guide for dummies." Unpublished manuscript.

Campos-Mercade, Pol, Armando Meier, Florian H Schneider, and Roberto A Weber. 2024. "What Money Shouldn't Buy: Aversion to Monetary Incentives for Health Behaviors." Working Paper.

Campos-Mercade, Pol, Armando N Meier, Florian H Schneider, and Erik Wengström. 2021a. "Prosociality predicts health behaviors during the COVID-19 pandemic." Journal of public economics, 195: 104367.

Campos-Mercade, Pol, Armando N Meier, Florian H Schneider, Stephan Meier, Devin Pope, and Erik Wengström. 2021b. "Monetary incentives increase COVID-19 vaccinations." Science, 374(6569): 879–882.

Deutchman, Paul. 2023. "People update their injunctive norm beliefs and moral judgements after receiving descriptive norm information."

Deutchman, Paul, Julia Marshall, Young-eun Lee, Felix Warneken, and Katherine McAuliffe. 2023. "Descriptive Norms Influence Children's Injunctive and Moral Norm Beliefs." Available at SSRN 4348267.

Dimant, Eugen, Michele Gelfand, Anna Hochleitner, and Silvia Sonderegger. 2024. "Strategic behavior with tight, loose, and polarized norms." Management Science.

Duch, Raymond, Edward Asiedu, Ryota Nakamura, Thomas Rouyard, Alberto Mayol, Adrian Barnett, Laurence Roope, Mara Violato, Dorcas Sowah, Piotr Kotlarz, *et al.* 2023. "Financial incentives for COVID-19 vaccines in a rural low-resource setting: a cluster-randomized trial." Nature Medicine, 29(12): 3193–3202.

Dwenger, Nadja, Henrik Kleven, Imran Rasul, and Johannes Rincke. 2016. "Extrinsic and intrinsic motivations for tax compliance: Evidence from a field experiment in Germany." American Economic Journal: Economic Policy, 8(3): 203–232.

Exley, Christine. 2018. "Incentives for prosocial behavior: The role of reputations." Management Science, 64(5): 2460–2471.

Gächter, Simon, Esther Kaiser, and Manfred Königstein. 2025. "Incentives crowd out voluntary cooperation: evidence from gift-exchange experiments." Experimental Economics, 1–32.

Gneezy, Uri, and Aldo Rustichini. 2000a. "A fine is a price." The journal of legal studies, 29(1): 1–17.

Gneezy, Uri, and Aldo Rustichini. 2000b. "Pay enough or don't pay at all." The Quarterly journal of economics, 115(3): 791–810.

Gneezy, Uri, *et al.* 2003. "The W effect of incentives." University of Chicago Graduate School of Business.

Holmås, Tor Helge, Egil Kjerstad, Hilde Lurås, and Odd Rune Straume. 2010. "Does monetary punishment crowd out pro-social motivation? A natural experiment on hospital length of stay." Journal of Economic Behavior & Organization, 75(2): 261–267.

Homonoff, Tatiana A. 2018. "Can small incentives have large effects? The impact of taxes versus bonuses on disposable bag use." American Economic Journal: Economic Policy, 10(4): 177–210.

Hossain, Tanjim, and John A List. 2012. "The behavioralist visits the factory: Increasing productivity using simple framing manipulations." Management Science, 58(12): 2151–2167.

Khazanov, Gabriela K, Rebecca Stewart, Matteo F Pieri, Candice Huang, Christopher T Robertson, K Aleks Schaefer, Hansoo Ko, and Jessica Fishman. 2023. "The effectiveness of financial incentives for COVID-19 vaccination: A systematic review." Preventive medicine, 172: 107538.

Klüver, Heike, Felix Hartmann, Macartan Humphreys, Ferdinand Geissler, and Johannes Giesecke. 2021. "Incentives can spur COVID-19 vaccination uptake." Proceedings of the National Academy of Sciences, 118(36): e2109543118.

Lane, Tom, Daniele Nosenzo, and Silvia Sonderegger. 2023. "Law and norms: Empirical evidence." American Economic Review, 113(5): 1255–1293.

Linardi, Sera, and Margaret A McConnell. 2008. "Volunteering and image concerns."

Mellström, Carl, and Magnus Johannesson. 2008. "Crowding out in blood donation: was Titmuss right?" Journal of the European Economic Association, 6(4): 845–863.

Muller, Robert. 2021. "Capital injection: Slovakia offers cash to over-60s to get Covid vaccine." Independent. https://web.archive.org/web/20230521171127/https://www.independent.co.uk/news/world/europe/slovakia-covid-vaccine-jab-cash-b1972963.html.

NHS. 2023. "Getting a COVID-19 vaccine." https://web.archive.org/web/20240119121750/https://www.nhs.

Panagopoulos, Costas. 2013. "Extrinsic rewards, intrinsic motivation and voting." The Journal of Politics, 75(1): 266–280.

Paravantes, Maria. 2022. "Fines Force Seniors in Greece to Rush and Get Vaccinated Against Covid-19." Greek Travel Pages (GTP). https://web.archive.org/web/20230602114057/https://news.gtp.gr/2022/01/18/fines-force-seniors-greece-rush-get-vaccinated-against-covid-19/.

Parmar, Shivani. 2021. "Harris County Will Pay You $100 To Get Your First Dose Of The COVID-19 Vaccine." Houston Public Media. https://web.archive.org/web/20231003035915/https://www.houstonpublicmedia.org/articles/news/health-science/2021/08/17/406205/harris-county-launches-100-incentive-program-to-combat-dwindling-vaccination-rate/.

Schneider, Florian H, Pol Campos-Mercade, Stephan Meier, Devin Pope, Erik Wengström, and Armando N Meier. 2023. "Financial incentives for vaccination do not have negative unintended consequences." Nature, 613(7944): 526–533.

Serra-Garcia, Marta, and Nora Szech. 2023. "Incentives and defaults can increase COVID-19 vaccine intentions and test demand." Management Science, 69(2): 1037–1049.

Sparkman, Gregg, Nathan Geiger, and Elke U Weber. 2022. "Americans experience a false social reality by underestimating popular climate policy support by nearly half." Nature communications, 13(1): 4779.

The Local Italy. 2022. "Over-50s in Italy without Covid booster face 100 euro fine." The Local. https://web.archive.org/web/20230608205613/https://www.thelocal.it/20220113/over-50s-in-italy-without-covid-booster-face-100-euro-fine.

Tversky, Amos, and Daniel Kahneman. 1979. "Prospect theory: an analysis of decision under risk." Econometrica, 47(2): 263–291.

Tversky, Amos, and Daniel Kahneman. 1991. "Loss aversion in riskless choice: A reference-dependent model." The quarterly journal of economics, 106(4): 1039–1061.

UK Health Security Agency. 2023. "Interactive map of cases." https://coronavirus.data.gov.uk/details/interactive-map/vaccinations.

UK Office for National Statistics. 2021. "Coronavirus and vaccine hesitancy, Great Britain: 9 August 2021." https://web.archive.org/web/20231023020401/https://www.ons.gov.uk/peoplepopulationandcommu

UK Office for National Statistics. 2023. "Average household income, UK: financial year ending 2022." https://web.archive.org/web/20240601072506/https://www.ons.gov.uk/peoplepopulationand

UK Office for National Statistics. 2024. "Population estimates for the UK, England, Wales, Scotland, and Northern Ireland: mid-2022." https://web.archive.org/web/20240606134223/https://www.ons.gov.uk/peoplepopulationandcommu

Wollbrant, Conny E, Mikael Knutsson, and Peter Martinsson. 2022. "Extrinsic rewards and crowding-out of prosocial behaviour." Nature Human Behaviour, 6(6): 774–781.

Appendix

A1.1 Demographic distribution

Table A1.1 provides a summary of the demographic distribution of our sample. The minimum age was 18 years old, and the maximum age was 88 years old.

Table A1.1: Summary statistics for demographic variables

	Summary
N	5,368
Gender	
Male	2,698 (50.3%)
Female	2,614 (48.7%)
Non-binary / third gender	34 (0.6%)
Prefer to self-describe	14 (0.3%)
Prefer not to say	8 (0.1%)
Income	
<20,000 GBP	1,499 (27.9%)
20,000-39,999 GBP	2,276 (42.4%)
40,000-59,999 GBP	1,058 (19.7%)
60,000-99,999 GBP	438 (8.2%)
More than 100,000 GBP	97 (1.8%)
Age	41.974 (13.540) 40

We report the frequency and percentage distribution for gender and income; and the mean, SD and median for age. For comparison, at the time of our study, 51% of the population was female in the UK, the median age was 40.7 years and the median household income was 32,300 GBP (UK Office for National Statistics 2024, 2023).

A1.1.1 Demographic distribution by treatment

Table A1.2: Summary statistics for demographic variables by incentive scheme

	No Incentive (N=1,061)	Large Penalty (N=1,083)	Large Reward (N=1,079)	Small Penalty (N=1,074)	Small Reward (N=1,071)	Total (N=5,368)
Gender						
Male	532 (50.1%)	538 (49.7%)	537 (49.8%)	546 (50.8%)	545 (50.9%)	2,698 (50.3%)
Female	518 (48.8%)	531 (49.0%)	529 (49.0%)	523 (48.7%)	513 (47.9%)	2,614 (48.7%)
Non-binary / third gender	8 (0.8%)	8 (0.7%)	9 (0.8%)	2 (0.2%)	7 (0.7%)	34 (0.6%)
Prefer to self-describe	1 (0.1%)	4 (0.4%)	2 (0.2%)	2 (0.2%)	5 (0.5%)	14 (0.3%)
Prefer not to say	2 (0.2%)	2 (0.2%)	2 (0.2%)	1 (0.1%)	1 (0.1%)	8 (0.1%)
Income						
<20,000 GBP	300 (28.3%)	309 (28.5%)	316 (29.3%)	290 (27.0%)	284 (26.5%)	1,499 (27.9%)
20,000-39,999 GBP	439 (41.4%)	456 (42.1%)	449 (41.6%)	465 (43.3%)	467 (43.6%)	2,276 (42.4%)
40,000-59,999 GBP	212 (20.0%)	212 (19.6%)	207 (19.2%)	215 (20.0%)	212 (19.8%)	1,058 (19.7%)
60,000-99,999 GBP	91 (8.6%)	82 (7.6%)	84 (7.8%)	84 (7.8%)	97 (9.1%)	438 (8.2%)
More than 100,000 GBP	19 (1.8%)	24 (2.2%)	23 (2.1%)	20 (1.9%)	11 (1.0%)	97 (1.8%)
Age	42.303 (13.876)	41.913 (13.479)	41.865 (13.235)	41.603 (13.190)	42.191 (13.920)	41.974 (13.540)

A1.2 Correlation of traits

Table A1.3: Correlation Table

	Trustworthiness	Honesty	Altruism
Trustworthiness	1.00		
Honesty	0.45***	1.00	
Altruism	0.44***	0.37***	1.00
Observations	5,368		

* p<0.05, ** p<0.01, *** p<0.001

A1.3 Distribution of esteem (S_k) ratings

As seen in Figures A1.1 and A1.2, attitudes towards COVID-19 vaccinations were not polarized in our sample -in which case the distribution of expected prosociality ratings would have followed an U shape (Dimant *et al.* 2024).

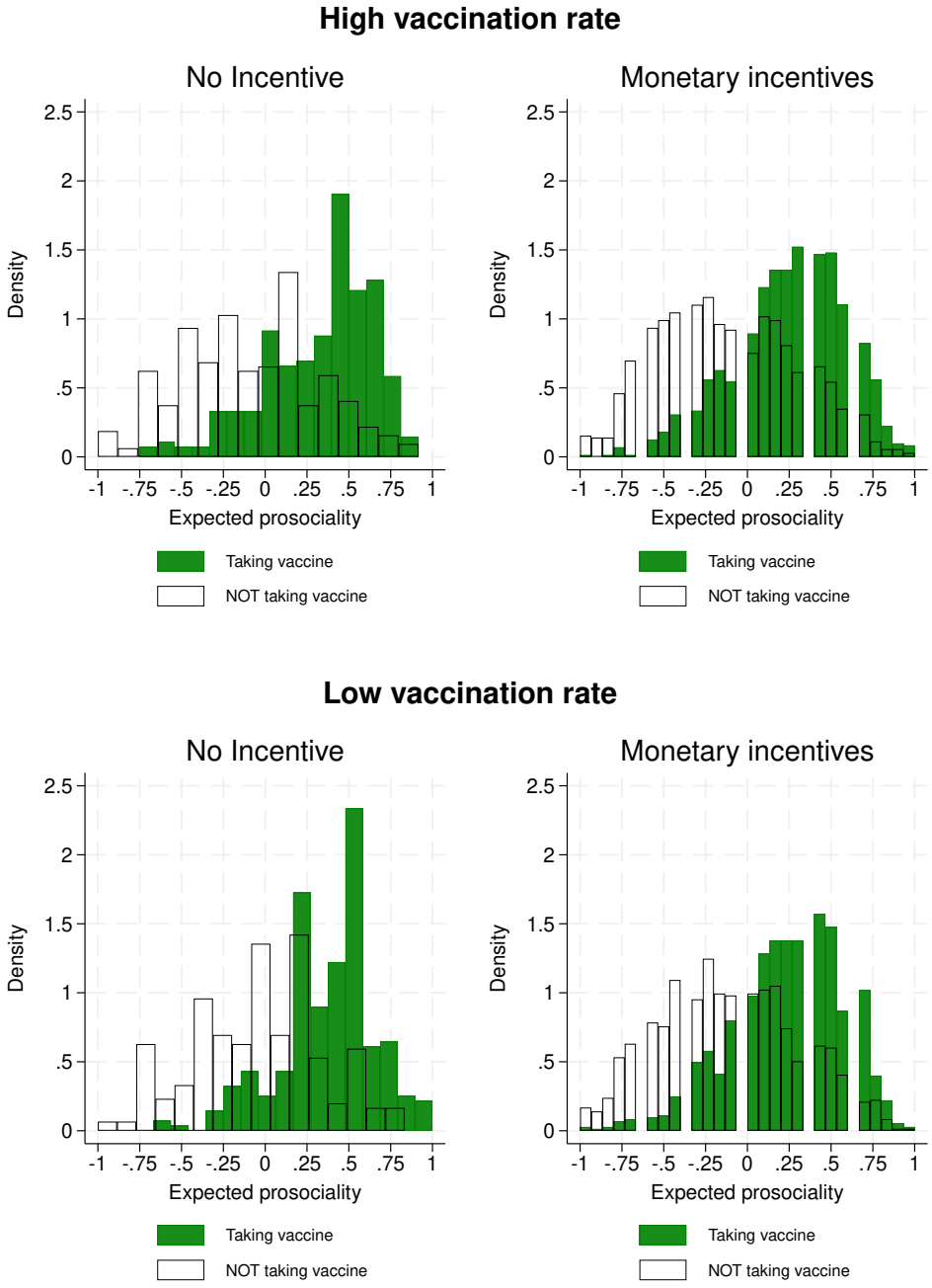


Figure A1.1: Expected prosociality ratings distribution, second-order beliefs

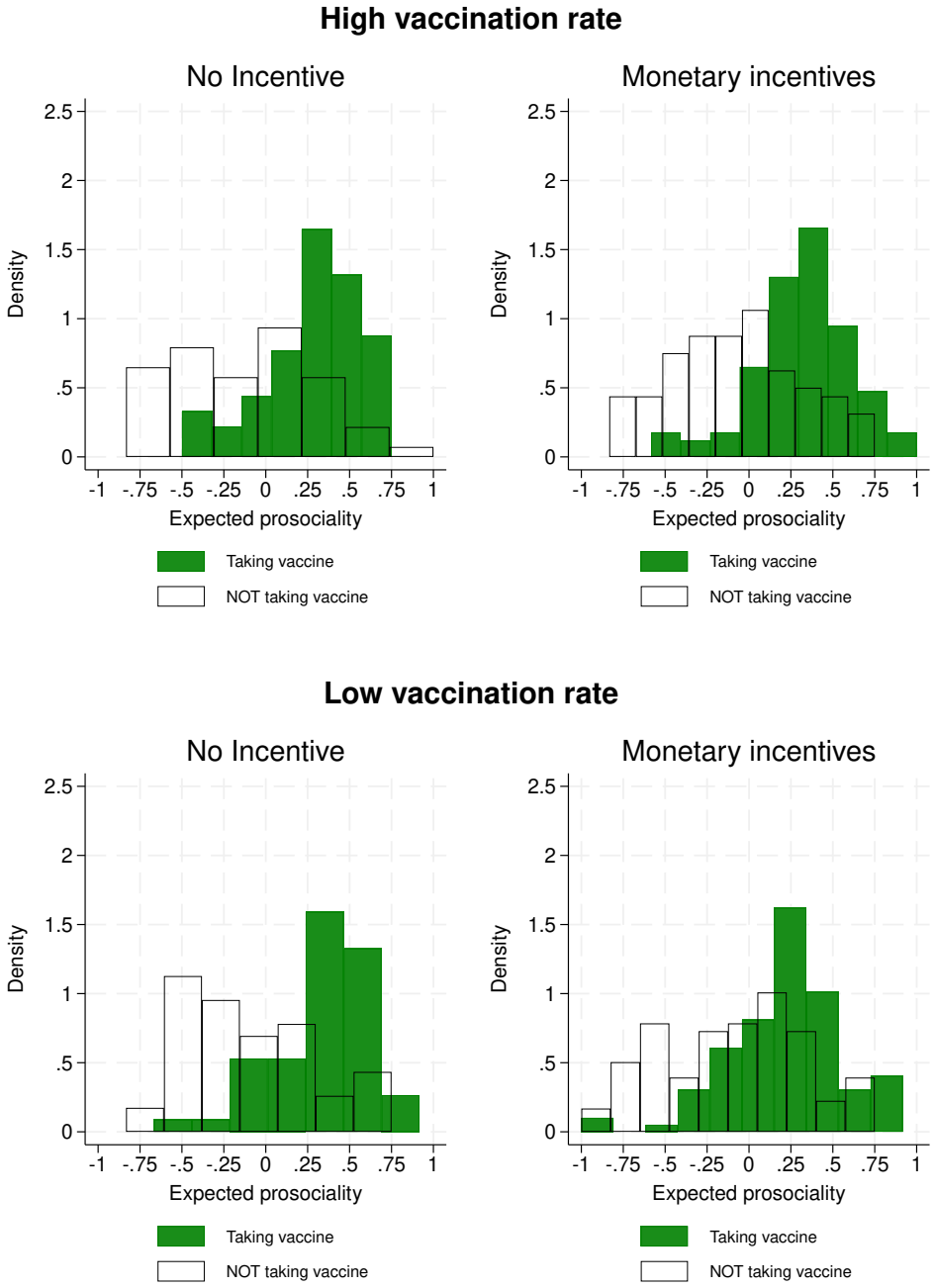


Figure A1.2: Expected prosociality ratings distribution, first-order beliefs

A1.4 Analyzing S_0 and S_1 independently

We used the Benjamini-Hochberg method to adjust the p-values of our tests of interest reported in this section.

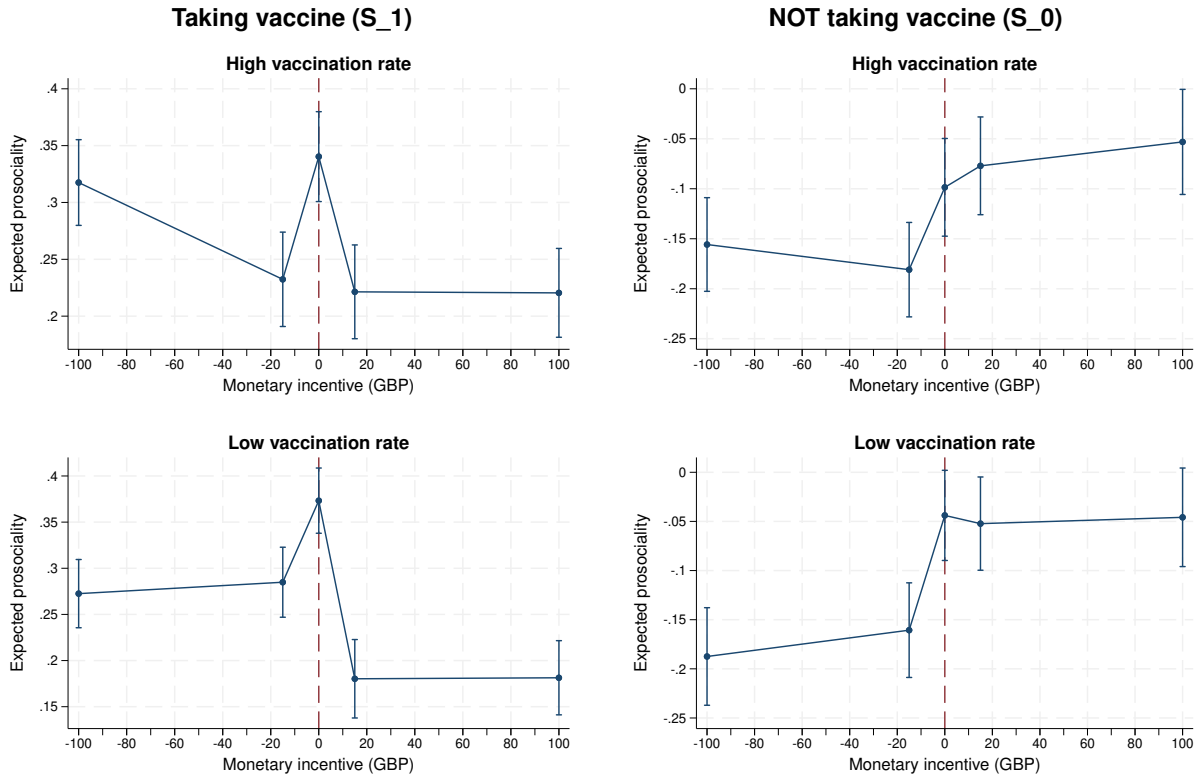


Figure A1.3: Esteem for taking (S_1) and not taking the vaccine (S_0) by treatment condition. CIs at the 95% confidence level, estimated by computing the standard errors of the means of S_1 and S_0 in each treatment condition.

A1.4.1 NOT taking vaccine (S_0)

As seen in Table A1.4, none of the monetary incentives changed S_0 significantly compared to baseline when vaccination rates were HIGH. When vaccination rates were LOW, S_0 remained unchanged for rewards but were significantly lower in the Penalty treatments.

According to our Wald tests, there was no difference between Large Penalty and Small Penalty, nor between Large Reward and Small Reward (all $p > 0.05$).

On the other hand, S_0 was significantly lower for Large Penalty compared to Large Reward ($p < 0.01$ for HIGH vaccination rate and $p < 0.01$ for LOW vaccination rate) and for Small Penalty compared to Small Reward (all $p < 0.01$).

In conclusion, we found that $S_0^b \geq S_0^j$ and $S_0^r > S_0^p$. These results are consistent with our theoretical framework.

Table A1.4: The effect of monetary incentives on expected prosociality when NOT taking the vaccine (S_0), by vaccination rate

	High Vaccination Rate (1)	Low Vaccination Rate (2)
Large Penalty	**-0.0478**	**-0.147*** **
Large Reward	**0.0572**	**-0.00637**
Small Penalty	**-0.0708**	**-0.123** **
Small Reward	**0.0363**	**-0.0121**
Constant	-0.324***	-0.200***
Controls	Yes	Yes
Observations	1343	1335

* p<0.05, ** p<0.01, *** p<0.001. Adjusted p-values in bold.

Control variables: gender, age, income (not reported)

A1.4.2 Taking vaccine (S_1)

Table A1.5: The effect of monetary incentives on expected prosociality when taking the vaccine (S_1), by vaccination rate

	High Vaccination Rate (1)	Low Vaccination Rate (2)
Large Penalty	**-0.0305**	**-0.0942** **
Large Reward	**-0.127*** **	**-0.186*** **
Small Penalty	**-0.113*** **	**-0.0773***
Small Reward	**-0.129*** **	**-0.190*** **
Constant	0.174***	0.212***
Controls	Yes	Yes
Observations	1335	1355

* p<0.05, ** p<0.01, *** p<0.001. Adjusted p-values in bold.

Control variables: gender, age, income (not reported)

As seen in Table A1.5, all monetary incentives lowered S_1 significantly compared to baseline, except for Large Penalty with HIGH vaccination rates.

According to our Wald tests, S_1 was significantly higher for Large Penalty compared to Small Penalty ($p < 0.01$) when vaccinations were HIGH. There was no such difference when vaccinations were LOW, and S_1 was no different between Large Reward and Small Reward (all $p > 0.05$).

Large Penalty had a significantly higher S_1 compared to Large Reward ($p < 0.01$ for both vaccination rates), but Small Penalty had a higher S_1 compared to Small Reward only when vaccinations were LOW ($p < 0.001$).

In summary, we found that $S_1^b \geq S_1^j$ and $S_1^p \geq S_1^r$. Additionally, we also found that $S_1^{large} \geq S_1^{small}$. All these results are consistent with our theoretical framework.

A1.5 Analyzing trustworthiness, honesty and altruism separately

Same as before, we used the Benjamini-Hochberg method to adjust the p-values of our tests of interest reported in this section.

A1.5.1 Net esteem from taking the COVID-19 vaccine

As shown in Tables A1.6, A1.7 and A1.8; the net esteem was positive and significant for all three traits ($p < 0.001$). Receiving the booster in baseline with HIGH and LOW vaccination rates increased perceptions of trustworthiness by 0.46 points (0.97 SD) and 0.37 points (0.79 SD); perceptions of honesty by 0.52 points (0.93 SD) and 0.5 points (0.9 SD); and perceptions of altruism by 0.37 points (0.66 SD) and 0.35 points (0.63 SD), respectively.

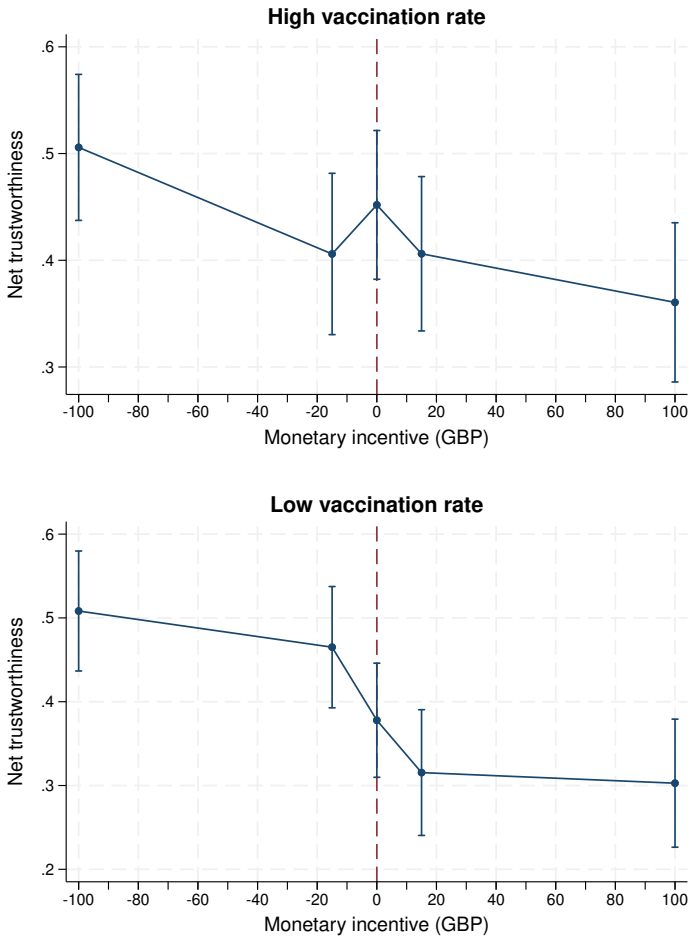


Figure A1.4: Net trustworthiness damage, with 95% CIs

We performed the same Wald tests as in the main analysis to check if net esteem remained positive in the presence of monetary incentives. We obtained $p < 0.001$ in all tests.

A1.5.2 Change in net esteem across different incentive schemes

The detrimental effect of rewards on social incentives were the strongest in honesty, and the weakest in trustworthiness.

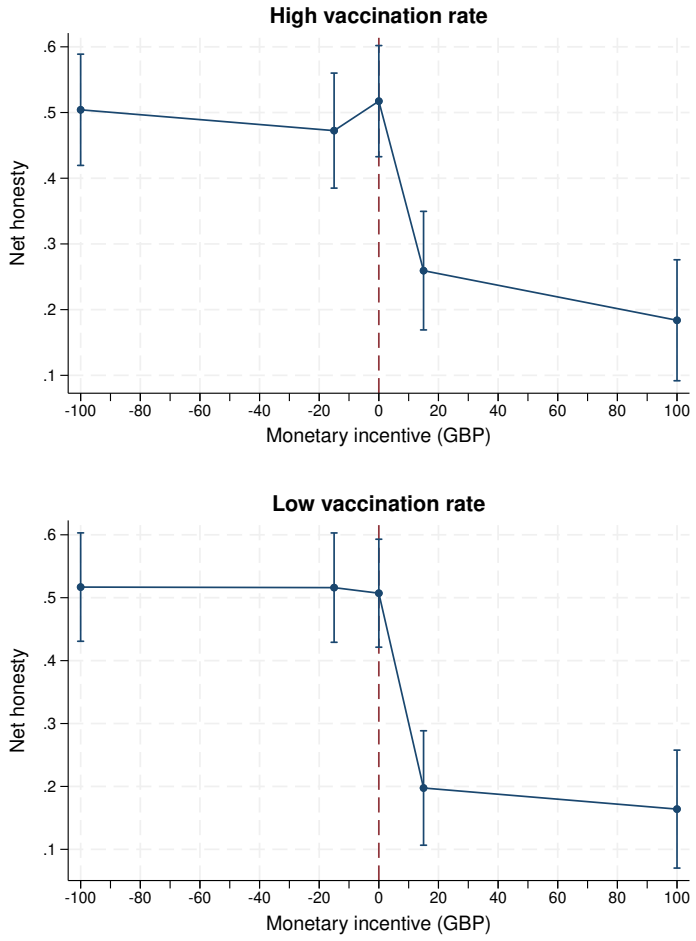


Figure A1.5: Net honesty damage, with 95% CIs

A1.5.2.1 Trustworthiness. Although rewards did not significantly damage the net trustworthiness, the direction of the effect was the same as in the main analysis. Interestingly, the Large Penalty treatment significantly *increased* net trustworthiness by 0.14 points (0.28 SD), albeit only in the LOW vaccination rate condition ($p = 0.02$). Joint Wald tests confirmed that Large Penalty was significantly different from Large Reward (HIGH vaccination rate: $p < 0.01$; LOW vaccination rate: $p < 0.001$), and that Small Penalty was significantly different from Small Reward ($p < 0.01$) for LOW vaccination rate only. On the other hand, the size of the incentives did not have a significant impact.

A1.5.2.2 Honesty. The effects of monetary incentives largely mirrored the effects found in the main analysis. Both reward treatments significantly decreased net honesty: Large Reward by 0.34 points (0.61 SD) for both vaccination rates; Small Reward by 0.27 (0.48 SD) and by 0.31 (0.55 SD) points for HIGH and LOW vaccination rates, respectively; all $p < 0.001$. Large Penalty was significantly different from Large Reward ($p < 0.001$ for both vaccination rates) and Small Penalty was significantly different from Small Reward (HIGH vaccination rate: $p = 0.001$; LOW vaccination rate: $p < 0.001$). Again, the size of the incentive did not make a difference.

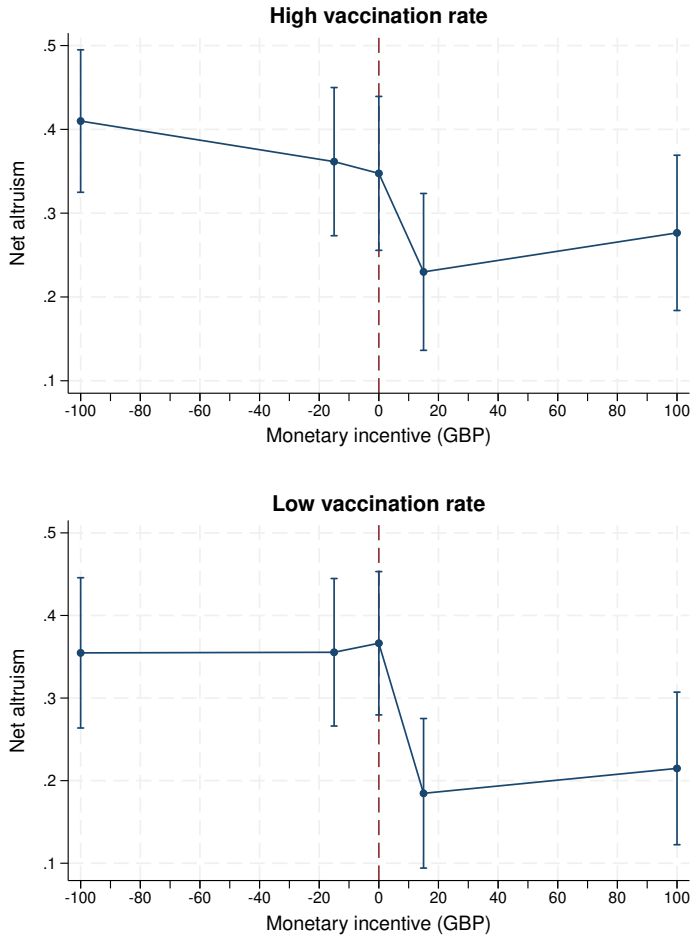


Figure A1.6: Net altruism damage, with 95% CIs

A1.5.2.3 Altruism. The Small Reward treatment significantly decreased net altruism by 0.16 points (0.28 SD; $p = 0.03$) for the HIGH vaccination rate and by 0.17 points (0.31 SD; $p = 0.01$) for the LOW vaccination rate. None of the other monetary incentives significantly reduced net altruism. Large Penalty was significantly different from Large Reward ($p < 0.05$) for LOW vaccination rate only. Small Penalty was significantly different from Small Reward for both vaccination rates (HIGH vaccination rate: $p < 0.05$; LOW vaccination rate: $p < 0.01$). Large and small incentives did not have a different impact.

Table A1.6: The effect of monetary incentives on trustworthiness, by vaccination rate

	High Vaccination Rate (1)	Low Vaccination Rate (2)
TakesVaccine=1	**0.462*** **	**0.374*** **
Large Penalty	-0.0513	-0.176***
Large Reward	0.0319	-0.0617
Small Penalty	-0.0199	-0.136***
Small Reward	-0.0158	-0.0516
TakesVaccine=1 × Large Penalty	**0.0428**	**0.134***
TakesVaccine=1 × Large Reward	**-0.105**	**-0.0715**
TakesVaccine=1 × Small Penalty	**-0.0589**	**0.0958**
TakesVaccine=1 × Small Reward	**-0.0630**	**-0.0589**
Constant	-0.0613	0.114**
Observations	2678	2690

* p<0.05, ** p<0.01, *** p<0.001. Adjusted p-values in bold.

Control variables: gender, age, income (not reported)

Table A1.7: The effect of monetary incentives on honesty, by vaccination rate

	High Vaccination Rate (1)	Low Vaccination Rate (2)
TakesVaccine=1	**0.524*** **	**0.501*** **
Large Penalty	-0.0215	-0.0952*
Large Reward	0.155***	0.118**
Small Penalty	-0.0727	-0.0811
Small Reward	0.122**	0.0891
TakesVaccine=1 × Large Penalty	**-0.0234**	**0.0164**
TakesVaccine=1 × Large Reward	**-0.341*** **	**-0.337*** **
TakesVaccine=1 × Small Penalty	**-0.0528**	**0.0179**
TakesVaccine=1 × Small Reward	**-0.271*** **	**-0.305*** **
Constant	-0.338***	-0.298***
Controls	Yes	Yes
Observations	2678	2690

* p<0.05, ** p<0.01, *** p<0.001. Adjusted p-values in bold.

Control variables: gender, age, income (not reported)

Table A1.8: The effect of monetary incentives on altruism, by vaccination rate

	High Vaccination Rate (1)	Low Vaccination Rate (2)
TakesVaccine=1	**0.372*** **	**0.354*** **
Large Penalty	-0.0703	-0.167***
Large Reward	-0.0193	-0.0704
Small Penalty	-0.120**	-0.148**
Small Reward	-0.0000677	-0.0696
TakesVaccine=1 × Large Penalty	**0.0314**	**0.00135**
TakesVaccine=1 × Large Reward	**-0.104**	**-0.137**
TakesVaccine=1 × Small Penalty	**-0.0170**	**0.0179**
TakesVaccine=1 × Small Reward	**-0.161***	**-0.173***
Constant	-0.501***	-0.411***
Controls	Yes	Yes
Observations	2678	2690

* p<0.05, ** p<0.01, *** p<0.001. Adjusted p-values in bold.

Control variables: gender, age, income (not reported)

A1.6 Exploratory findings: rewards increase variance in S_1 ratings, but only when vaccination rate is low

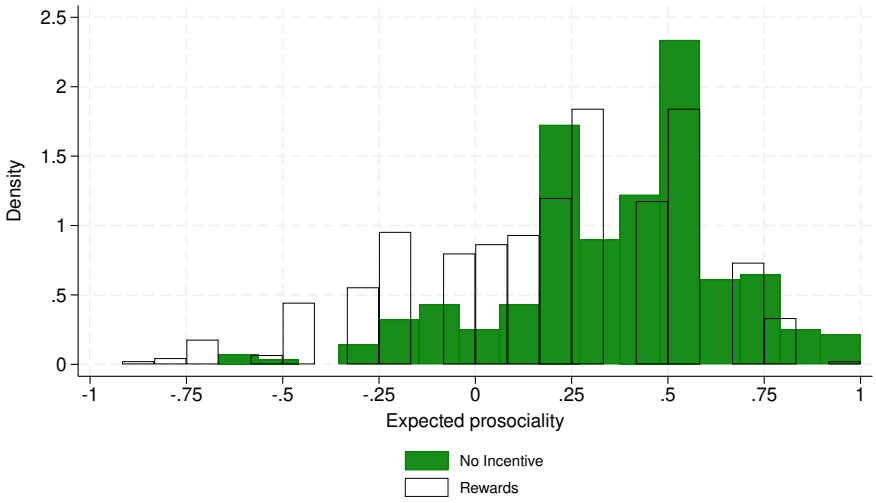
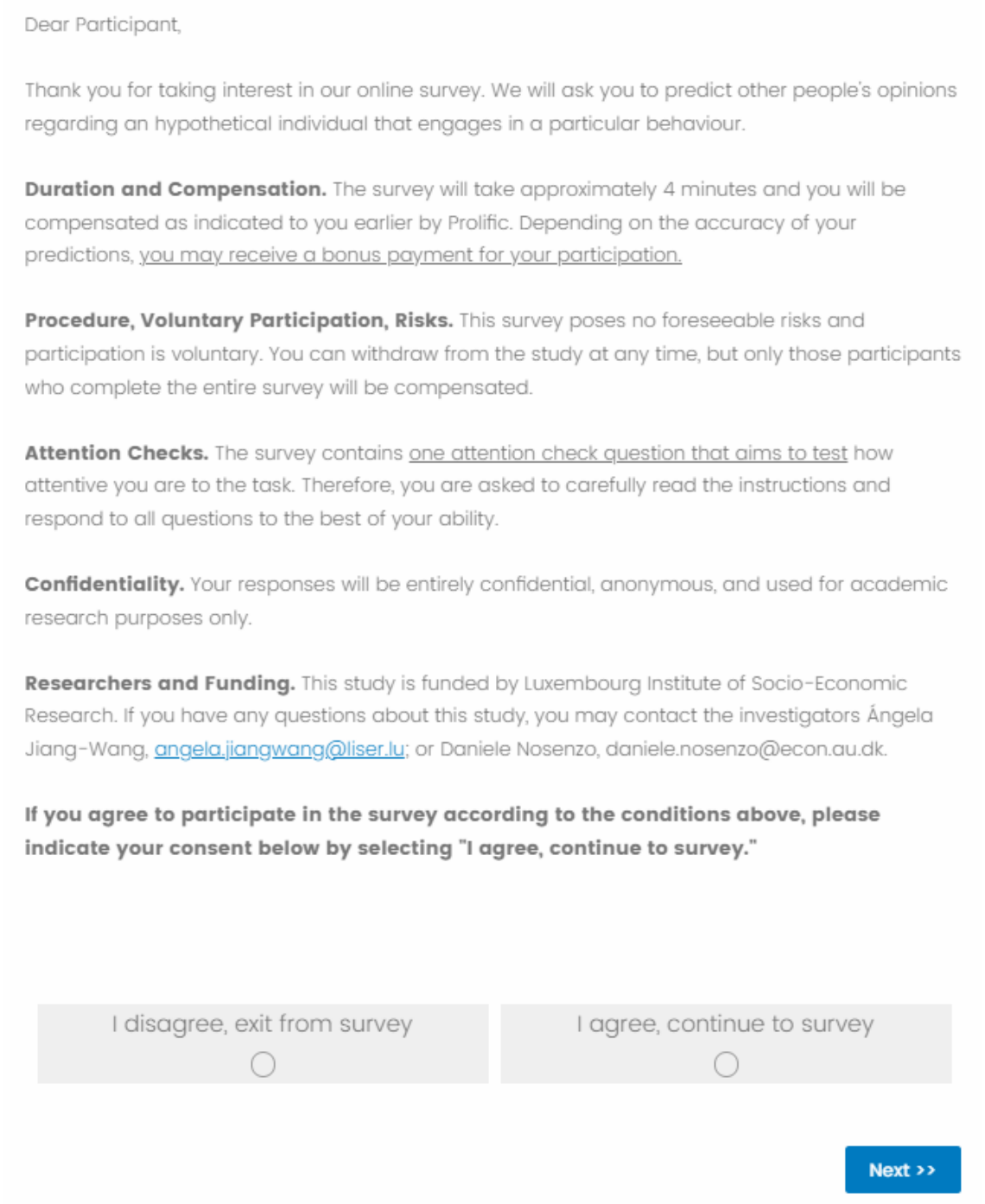


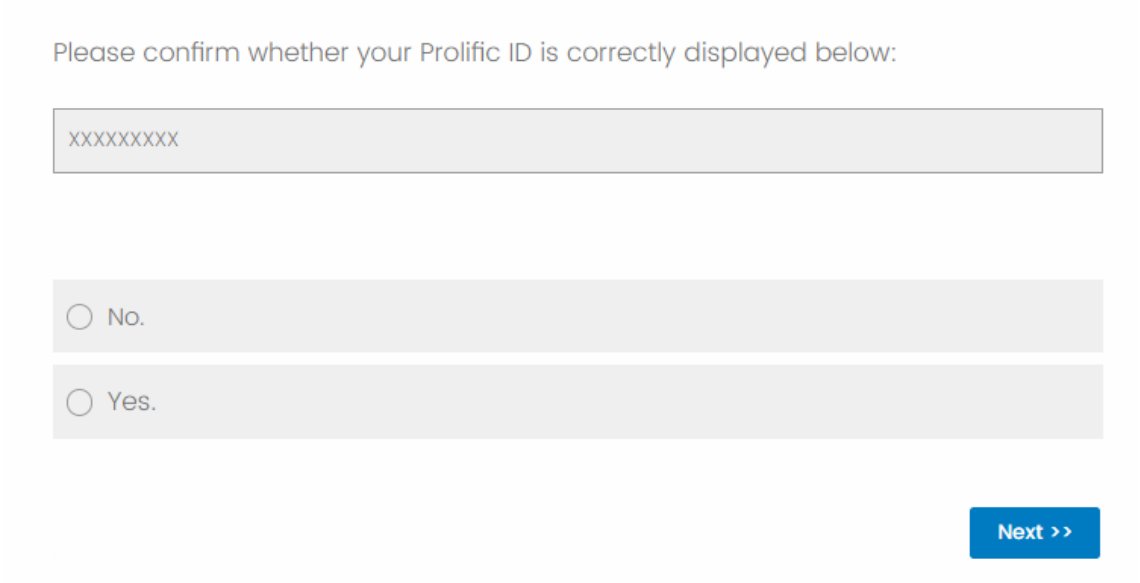
Figure A1.7: S_1 RATINGS, LOW VACCINATION RATE

A Levene's test revealed that the variance of the S_1 ratings was not equal between the different incentive schemes only when vaccination rate was low ($p < 0.01$). Specifically, we found that large and small rewards increased the variance of S_1 ratings compared to baseline ($p < 0.05$ and $p < 0.001$, respectively).

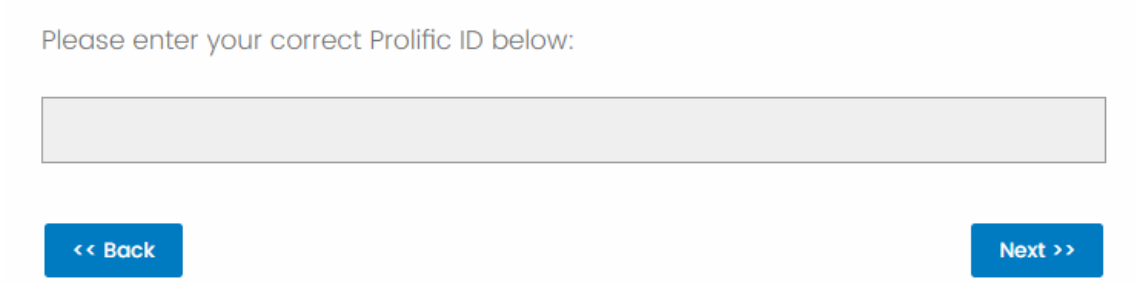
A1.7 Experiment instructions: second-order beliefs

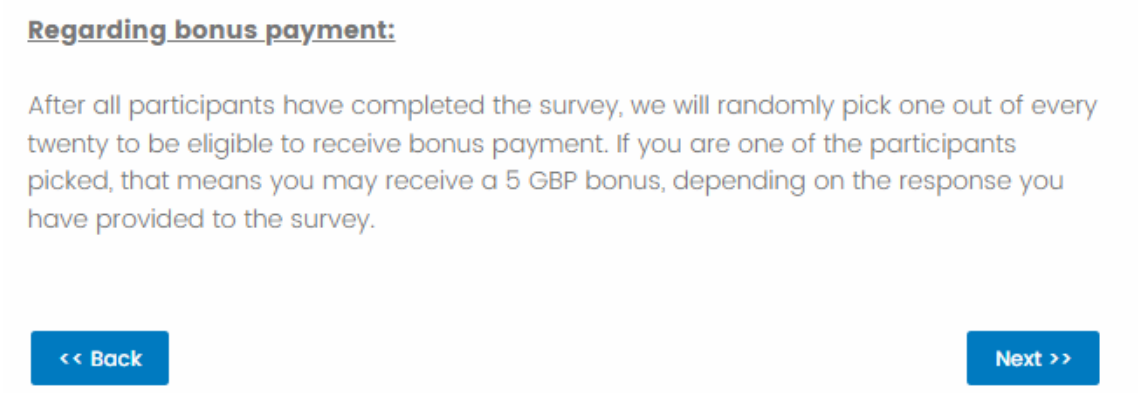
We present screenshots of our main experiment eliciting second-order beliefs. All questions were compulsory except for the attention check and the final "comment" box.





In case participants selected "no", then the following screen would appear:





Information about this survey:

In this survey your task is to guess the most common answers given to questions in a previous survey.

This previous survey described a hypothetical person's behaviour in a fictitious situation, and we asked participants their opinions about how this person might have behaved in other situations. We listed possible behaviours in these other situations and asked respondents how likely it is that this person might have done them in the recent past. For each question, there were eight possible responses, as shown below, of which respondents had to select exactly one.

All participants in the previous survey were located in the UK and recruited online on Prolific.

After you have completed this survey, we will randomly select one of the questions in which we asked you to predict participants' answers in the previous survey. If you correctly predicted the most common answer to this question, and if you are one of the participants selected as eligible for bonus payment, we will give you a 5 GBP bonus.

<< Back Next >>

We will now go through an example of a possible set of questions and illustrate how you would respond to them.

<< Back Next >>

Example questions

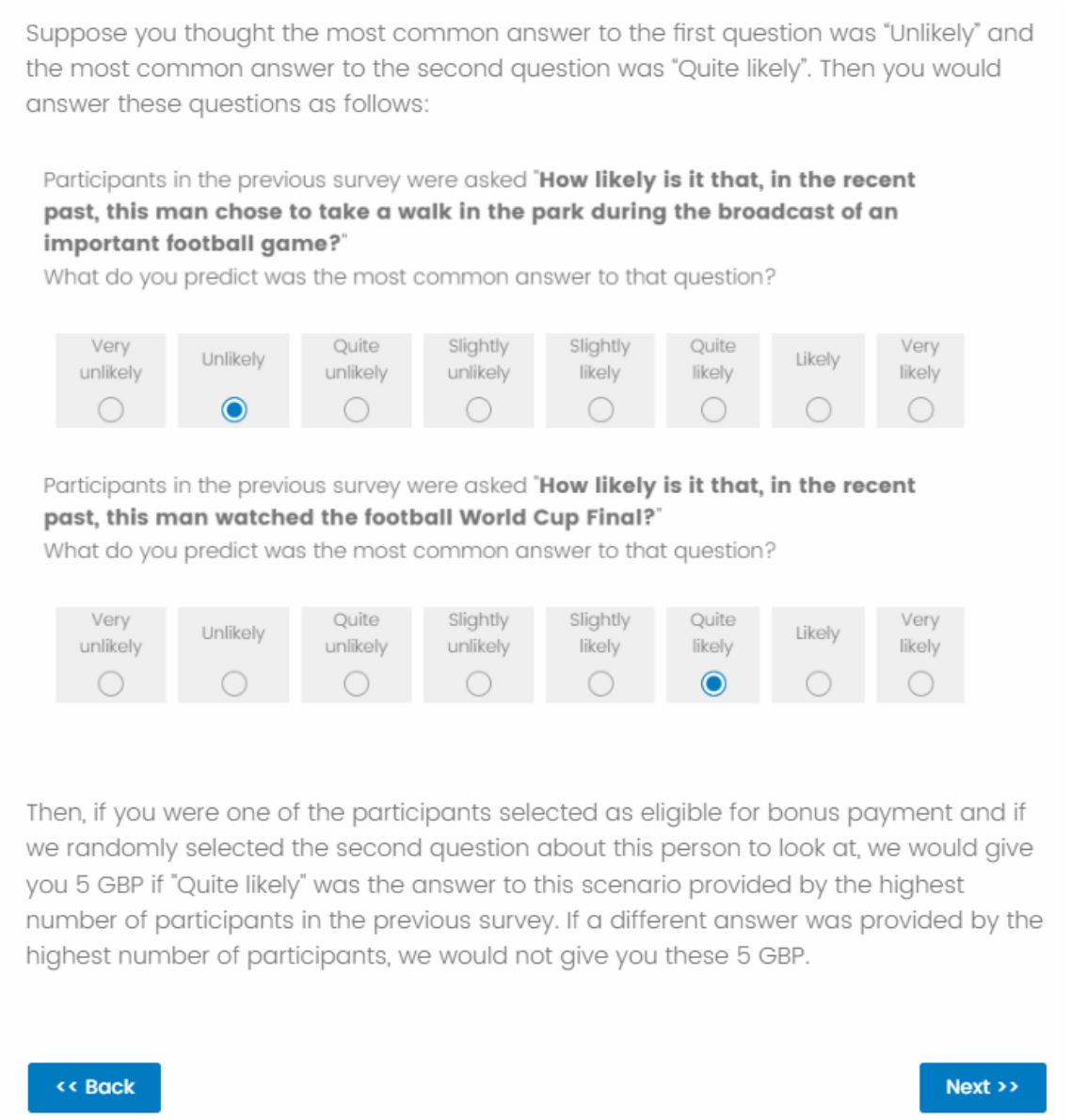
Suppose that we presented participants of the previous survey with the following hypothetical person's behaviour:

A man is planning to attend a friend's wedding on Saturday. The man is a big football fan and, two days before the wedding, he is offered free tickets to watch an important football match. The man decides to take the tickets. On the Saturday, he goes to the football match, and tells his friend he is too ill to attend the wedding.

Suppose participants were asked:

(1) "How likely is it that, in the recent past, this man chose to take a walk in the park during the broadcast of an important football game?"

(2) "How likely is it that, in the recent past, this man watched the football World Cup Final?"


Suppose you thought the most common answer to the first question was "Unlikely" and the most common answer to the second question was "Quite likely". Then you would answer these questions as follows:
Participants in the previous survey were asked "How likely is it that, in the recent past, this man chose to take a walk in the park during the broadcast of an important football game?"
What do you predict was the most common answer to that question?
Very unlikely
Unlikely
Quite unlikely
Slightly unlikely
Slightly likely
Quite likely
Likely
Very likely
Participants in the previous survey were asked "How likely is it that, in the recent past, this man watched the football World Cup Final?"
What do you predict was the most common answer to that question?
Very unlikely
Unlikely
Quite unlikely
Slightly unlikely
Slightly likely
Quite likely
Likely
Very likely
Then, if you were one of the participants selected as eligible for bonus payment and if we randomly selected the second question about this person to look at, we would give you 5 GBP if "Quite likely" was the answer to this scenario provided by the highest number of participants in the previous survey. If a different answer was provided by the highest number of participants, we would not give you these 5 GBP.
<< Back
Next >>

Now you are about to start the task. This is your last opportunity to go back and read the instructions. Please press the "Next >>" button to start the task.

This is a question for us to check that you are paying attention to the survey. Please let us know that you are reading all our questions carefully by selecting "Very unlikely" on the response scale

Very unlikely	Unlikely	Quite unlikely	Slightly unlikely	Slightly likely	Quite likely	Likely	Very likely
○	○	○	○	○	○	○	○

<< Back Next >>

Participants in the previous survey were presented with the following hypothetical person's behaviour:

A man works as a janitor in a residential home for the elderly in a medium-sized UK town. For this reason, he is eligible to receive the seasonal COVID-19 booster vaccine.

One day, the man receives a letter from the local Council inviting him to get the seasonal booster vaccine within the next two weeks. The letter also says that 15% of eligible Council residents (that is, residents at increased risk or in close contact with people at increased risk) took the booster in the previous season.

The man does NOT take the seasonal COVID-19 booster within the following two weeks.

Please answer the following questions about how participants in the previous survey said this man might have behaved in the recent past.

Participants in the previous survey were asked:

"**How likely is it that this man read at least two books per month?**"
What do you predict was the most common answer to that question?

Very unlikely	Unlikely	Quite unlikely	Slightly unlikely	Slightly likely	Quite likely	Likely	Very likely
◯	◯	◯	◯	◯	◯	◯	◯

"**How likely is it that this man kept a healthy diet, avoiding fatty foods and refined sugar?**"
What do you predict was the most common answer to that question?

Very unlikely	Unlikely	Quite unlikely	Slightly unlikely	Slightly likely	Quite likely	Likely	Very likely
◯	◯	◯	◯	◯	◯	◯	◯

"**How likely is it that this man made a donation to a friend who ran the TCS London marathon for charity?**"
What do you predict was the most common answer to that question?

Very unlikely	Unlikely	Quite unlikely	Slightly unlikely	Slightly likely	Quite likely	Likely	Very likely
◯	◯	◯	◯	◯	◯	◯	◯

"**How likely is it that this man kept fit by regularly going to the gym?**"
What do you predict was the most common answer to that question?

Very unlikely	Unlikely	Quite unlikely	Slightly unlikely	Slightly likely	Quite likely	Likely	Very likely
○	○	○	○	○	○	○	○

"How likely is it that this man kept a promise made to a friend?"
What do you predict was the most common answer to that question?

Very unlikely	Unlikely	Quite unlikely	Slightly unlikely	Slightly likely	Quite likely	Likely	Very likely
○	○	○	○	○	○	○	○

"**If a cashier accidentally gave this man more change than he was due, how likely is it that he returned the extra change?**"
What do you predict was the most common answer to that question?

Very unlikely	Unlikely	Quite unlikely	Slightly unlikely	Slightly likely	Quite likely	Likely	Very likely
○	○	○	○	○	○	○	○

Remember that, if one of these questions is selected for payment, you can earn 5 GBP from it only if you correctly predict the most common response to the question in the previous survey. Remember that all participants in the previous survey were located in the UK and recruited online on Prolific.

Next >>

How do you describe yourself?

- ◯ Male
- ◯ Female
- ◯ Non-binary / third gender
- ◯ Prefer to self-describe
- ◯ Prefer not to say

<< Back Next >>

What is your age? (please type the number)

<< Back Next >>

What is your annual personal income before taxes?

- ◯ Less than 20,000 GBP
- ◯ 20,000-39,999 GBP
- ◯ 40,000-59,999 GBP
- ◯ 60,000-99,999 GBP
- ◯ More than 100,000 GBP
- ◯ I prefer not to say

For comments or remarks on the survey, please use the space provided below.

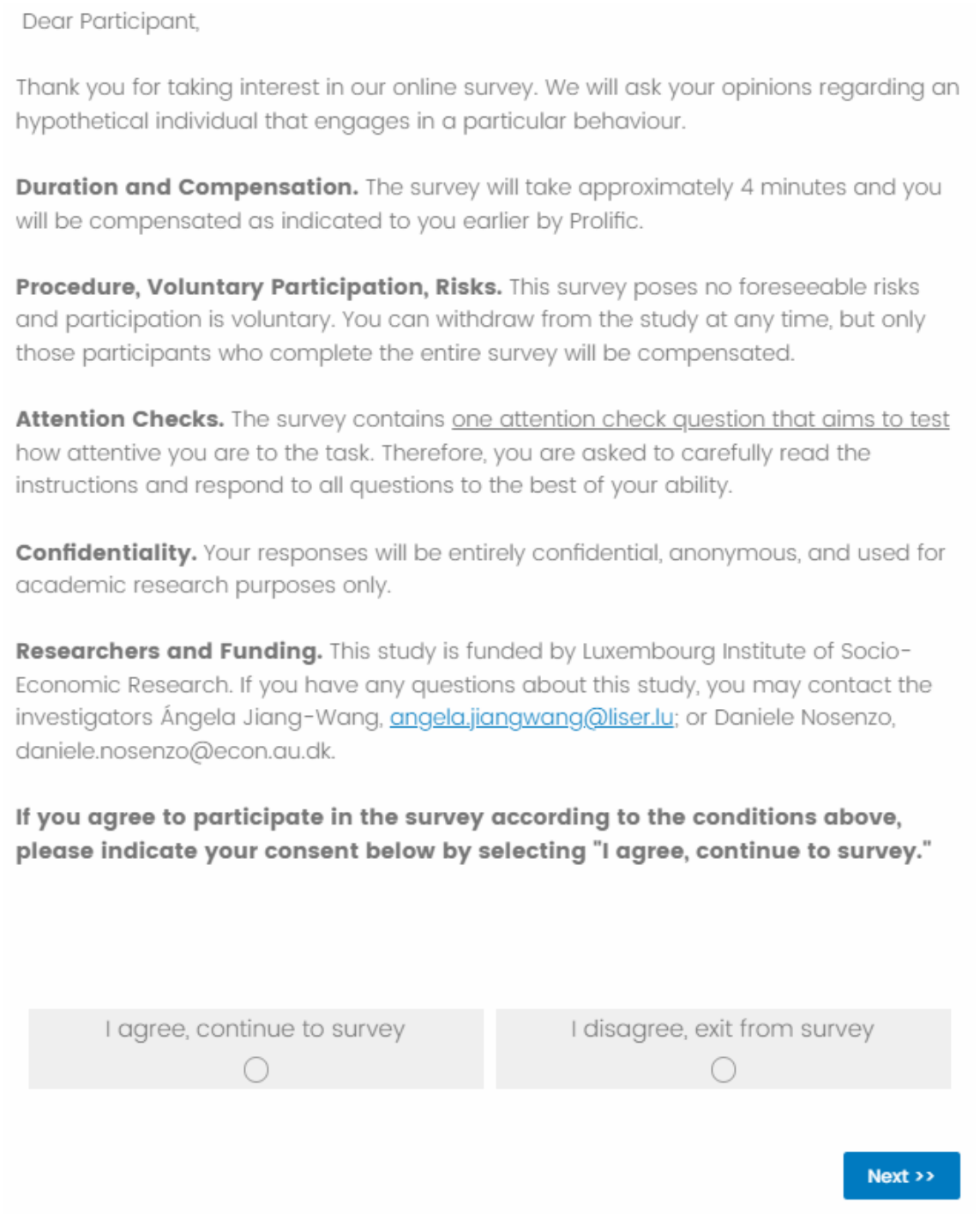
<< PLEASE PRESS THE "NEXT >>" BUTTON BELOW TO SAVE YOUR ENTRIES AND RECEIVE YOUR COMPLETION CODE>>

Thank you very much for participating in this study. Your participation was invaluable to us.

<< Back Next >>

A1.8 Experiment instructions: first-order beliefs

We present screenshots of the first-order beliefs experiment. All questions were compulsory except for the attention check and the final "comment" box. The questions about the Prolific ID, demographics and the end of the experiment were identical to the ones in our main experiment, hence we are not showing them again.



Information about this survey:

This survey will describe a hypothetical person's behaviour in a fictitious situation, and we will ask your opinion about how this person might have behaved in other situations. We will list possible behaviours in these other situations and we will ask you how likely it is that this person might have done them in the recent past. For each question there will be eight possible responses, as shown below, of which you must select exactly one.

Very unlikely	Unlikely	Quite unlikely	Slightly unlikely	Slightly likely	Quite likely	Likely	Very likely
○	○	○	○	○	○	○	○

<< Back Next >>

We will now go through an example of a possible set of questions and illustrate how you would respond to them.

<< Back Next >>

An example of a person's behaviour

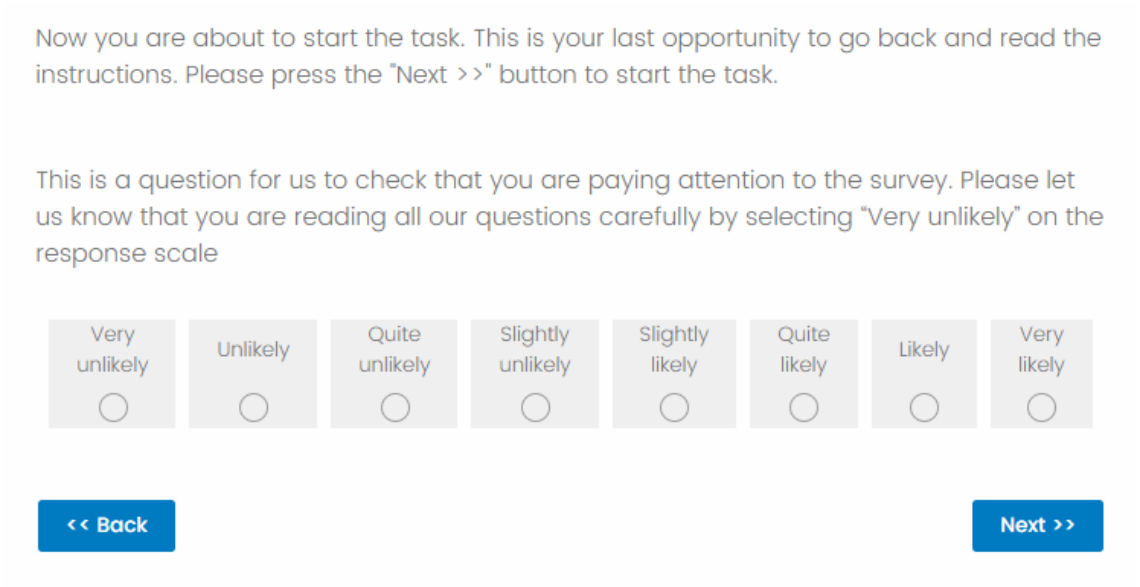
A man is planning to attend a friend's wedding on Saturday. The man is a big football fan and, two days before the wedding, he is offered free tickets to watch an important football match. The man decides to take the tickets. On the Saturday, he goes to the football match, and tells his friend he is too ill to attend the wedding.

Suppose we asked you:

(1) "How likely is it that, in the recent past, this man chose to take a walk in the park during the broadcast of an important football game?"

(2) "How likely is it that, in the recent past, this man watched the football World Cup Final?"

Suppose you thought the answer to the first question was "Unlikely" and the answer to the second question was "Quite likely". Then you would answer these questions as follows:



A man works as a janitor in a residential home for the elderly in a medium-sized UK town. For this reason, he is eligible to receive the seasonal COVID-19 booster vaccine.

One day, the man receives a letter from the local Council inviting him to get the seasonal booster vaccine within the next two weeks. The letter also says that 84% of eligible Council residents (that is, residents at increased risk or in close contact with people at increased risk) took the booster in the previous season.

The letter further states that, this year, the Council has introduced a penalty of 15 GBP for all eligible residents who do not take the booster within the next two weeks. If their vaccination is not confirmed in the NHS registry, they will receive a penalty notice at home.

The man takes the seasonal COVID-19 booster within the following two weeks.

Please answer the following questions about how this man might have behaved in the recent past:

How likely is it that this man kept fit by regularly going to the gym?

Very unlikely	Unlikely	Quite unlikely	Slightly unlikely	Slightly likely	Quite likely	Likely	Very likely
○	○	○	○	○	○	○	○

How likely is it that this man made a donation to a friend who ran the TCS London marathon for charity?

Very unlikely	Unlikely	Quite unlikely	Slightly unlikely	Slightly likely	Quite likely	Likely	Very likely
○	○	○	○	○	○	○	○

How likely is it that this man kept a healthy diet, avoiding fatty foods and refined sugar?

Very unlikely	Unlikely	Quite unlikely	Slightly unlikely	Slightly likely	Quite likely	Likely	Very likely
○	○	○	○	○	○	○	○

If a cashier accidentally gave this man more change than he was due, how likely is it that he returned the extra change?

Very unlikely	Unlikely	Quite unlikely	Slightly unlikely	Slightly likely	Quite likely	Likely	Very likely
○	○	○	○	○	○	○	○

How likely is it that this man read at least two books per month?

Very unlikely	Unlikely	Quite unlikely	Slightly unlikely	Slightly likely	Quite likely	Likely	Very likely
○	○	○	○	○	○	○	○

How likely is it that this man kept a promise made to a friend?

Very unlikely	Unlikely	Quite unlikely	Slightly unlikely	Slightly likely	Quite likely	Likely	Very likely
○	○	○	○	○	○	○	○

Next >>

Chapter 2

You might be underestimating how sustainable others are

You might be underestimating how sustainable others are

with Francesco Fallucchi (University of Bergamo), Philippe Van Kerm (University of Luxembourg), and Bertrand Verheyden (Luxembourg Institute of Socio-Economic Research)

2.1 Introduction

Climate change is a serious global concern. Yet, around the world, people tend to underestimate how much others care about addressing it (Andre *et al.* 2024a). This underestimation extends to others' concern about climate change (Geiger and Swim 2016, Sparkman, Geiger and Weber 2022, Leviston, Walker and Morwinski 2013), willingness to contribute money to fight against it (Andre *et al.* 2024a,b), and support for climate policies (Sparkman, Geiger and Weber 2022, Sokoloski, Markowitz and Bidwell 2018). As a result, there is a systematic misperception of the social norms surrounding climate change mitigation.

Our expectations about others matter: both *empirical expectations* (beliefs about what others do) and *normative expectations* (beliefs about what others deem appropriate to do) significantly influence our own behavior (Bicchieri 2005, 2016). In other words, many of our preferences are conditional, and this holds true for pro-environmental behaviors as well (see Cialdini and Jacobson 2021, Saracevic and Schlegelmilch 2021, for recent reviews on the topic). Therefore, misperceived social norms can actively hinder climate change mitigation efforts, making it essential for scientists and practitioners to understand and address them (Geiger and Swim 2016, Andre *et al.* 2024b, Clayton *et al.* 2015).

While many studies on misperceptions focus on general concerns and attitudes about climate change, very few examine misperceptions related to everyday sustainable behaviors. Similarly, research on misperceptions of policy support typically investigates policies that target companies or public institutions and benefit individuals. However, to the extent of our knowledge, no studies have examined policies that target, and potentially constrain, individual behavior. In this paper, we examine whether Luxembourgers underestimate how much others' engage in everyday sustainable behaviors, their personal norms regarding these behaviors (which is the behavior level they deem appropriate), and their support for policies restricting unsustainable behaviors. We are interested in investigating these misperceptions for three main reasons: 1) Individual behavioral change is essential to reduce global emissions and remains a key target for effective climate action (Dubois *et al.* 2019); 2) While opinions, attitudes, and willingness to donate money can be private and unobservable,

especially if people choose not to openly express them or discuss them, behaviors are typically public and visible, making them harder to misperceive. Do individuals still misperceive behaviors that are observable on a daily basis, and the personal norms associated with them?; 3) Policies that restrict unsustainable behaviors often require individuals to make personal sacrifices, making support for such policies more costly and less popular (Swim and Geiger 2021). Do individuals still misperceive support for these more personally demanding policies?

Household consumption accounts for nearly 75% of global carbon emissions (Druckman and Jackson 2016, Dubois *et al.* 2019, Hertwich and Peters 2009), particularly in the areas of food, housing, and transportation. There is broad consensus that systematic and sustained behavioral change, along with climate policies targeting household consumption, can lead to substantial emission reductions, especially in high-income countries. For example, switching to a vegetarian diet, reducing home heating needs, and using public transport instead of a car for commuting can save up to 4.8 tonnes of CO_2eq per year, about 37% of the 13 tonnes of CO_2eq emitted annually by the average Luxembourger (Hitaj, Igos and Gibon 2022). Hence, it is essential for us to understand the barriers that currently hinder behavioral change, such as potential misperceptions surrounding these behaviors and the policies addressing them (Clayton *et al.* 2015). In the remainder of the paper, we use *sustainable behaviors* as an umbrella term to refer to both carbon-saving and pro-environmental behaviors, and *unsustainable behaviors* to denote high-carbon behaviors.

We conducted a well-powered, pre-registered online survey (N = 1,292) among Luxembourgish residents and cross-border workers to examine whether people significantly underestimate the extent to which others engage in the three behaviors with the highest carbon-emission reduction potential: consuming vegetarian meals, reducing home heating, and using public transportation (Hitaj, Igos and Gibon 2022). We also examined whether they significantly underestimate the level of engagement others consider appropriate for these behaviors, as well as their support for policies that restrict and tax the associated unsustainable behaviors within these domains. Participants reported their actual engagement levels in these three behaviors, which served as our main measure of behavior. Additionally, we also investigated whether people underestimate how much money others donate to reduce carbon emissions. Participants could choose whether to donate a portion of a €250 lottery prize to a carbon-offset project. This allowed us to obtain an incentivized behavioral measure to complement the self-reported data and to test whether we could replicate the findings of Andre *et al.* (2024b) in our Luxembourgish sample. Subsequently, participants reported what they believed to be the ethically appropriate engagement levels for the previously mentioned behaviors from a sustainability perspective, which served as our measure of personal norms. Finally, they

indicated whether they would support six hypothetical policies that restrict and tax the associated unsustainable behaviors in these three behavioral domains (animal protein consumption, home heating, and car usage). This was our measure of policy support.

We elicited social expectations through incentivized second-order beliefs. That is, we asked participants to guess the most common responses given by all survey participants regarding behaviors and personal norms, which served as our measures of empirical and normative expectations, respectively. Participants also estimated, out of 100, how many other respondents supported each policy, which served as our measure of expected policy support. Social expectations were elicited in an incentive-compatible manner: participants scored points for accurate guesses and could earn a bonus of up to €30 if they ranked among the top 30% of scorers at the end of the survey.

A key contribution of our study is that we examine the existence of these underestimations not only at the extensive margin but also at the intensive margin, that is, how much effort others put into engaging in sustainable practices. For instance, even if people believe that most others consume vegetarian meals once in a while, they may still underestimate how often others actually do so. To the best of our knowledge, only one other study has looked at underestimations of behaviors and personal norms related to everyday sustainable practices (Andre *et al.* 2024b). The authors asked participants to estimate how many others engaged in behaviors such as restricting meat consumption or regularly using environmentally friendly alternatives to fossil-fueled cars, as well as how many others believed these behaviors ought to be done. Our study builds on and extends this work by capturing not only potential misperceptions about how many people engage in sustainable behaviors, but also the intensity with which they do so. Even when individuals aim to adopt more sustainable practices, they may lack a clear reference point for what constitutes an appropriate frequency or intensity. Thus, even if they have accurate perceptions of prevalence, underestimating the intensity of others' actions may still lead them to align with a misperceived social norm.

A second advantage of asking about behaviors and expectations at the intensive margin is that it reduces variability in interpretation. What constitutes "regular behavior" can vary between participants. Self-image motivations may also lead individuals to interpret these terms favorably, believing they engage in sustainable behaviors often even if their actual effort is minimal (Dana, Weber and Kuang 2007). Similarly, this variability in interpretation can also bias one's beliefs about others. In contrast, it is harder to distort beliefs about precise, objective measures of quantity or frequency. Since objective and precise measures of behavior are harder to misperceive, we can test whether misperceptions are robust to these more concrete indicators.

In line with our pre-registered hypotheses, we document systematic and widespread underes-

timation of others' sustainable behaviors, personal norms, and policy support. The magnitude of these underestimations is significant across all behavioral domains, with particularly pronounced effects in vegetarian consumption and home heating reduction. We also replicate the findings of Andre *et al.* (2024b), observing that participants also significantly underestimate how much others donate. However, the size of this misperception is smaller in our study; in Section 2.3.4, we explore potential explanations for this discrepancy.

Contrary to Andre *et al.* (2024b), we find that underestimations are larger for behaviors (mean Cohen's d = 0.34) than for personal norms (mean Cohen's d = 0.20). Whereas 60% and 62% of participants underestimated the intensity of others' vegetarian consumption and home heating reduction, respectively, 52% and 57% underestimated the corresponding personal norms. This pattern was reversed in the domain of public mobility: while only 42% of participants underestimated the behavioral intensity, 60% underestimated the intensity of personal norms. On average, 56% of participants in our sample underestimated others' sustainable behaviors and personal norms. The proportion of underestimators in our study is lower than that reported by Andre *et al.* (2024b) regarding willingness to fight climate change, which is expected given that our measures of behavior and norms are more observable and therefore less prone to misperception.

Interestingly, we also find that the magnitude of underestimations is larger for policy support than for behaviors and personal norms (mean Cohen's d = 0.43). On average, 80% of participants underestimated the number of supporters of a policy. Moreover, while participants believed that none of the six policies would be supported by a majority, in reality, four of them were. Although underestimations are significant across all policies, their magnitude is smaller than those reported in previous studies about misperceptions in climate policies (Sparkman, Geiger and Weber 2022), which is expected given that the policies we examined are more personally costly to support. Consistent with this, both the overall prevalence of support and the estimated prevalence are lower than those found by Sparkman, Geiger and Weber (2022). Interestingly, however, the share of underestimators in our sample is similar to theirs.

We were also interested in exploring the mechanisms underlying these underestimations. Specifically, we aimed to assess whether our results may be driven by pluralistic ignorance and/or the false consensus effect, two mechanisms commonly discussed in the literature as sources of misperceptions regarding pro-environmental social norms. To this end, we conducted heterogeneity analyses by examining underestimations separately for individuals who reported levels of sustainable behavior above the mean (who we refer to as "the sustainable") and those who reported levels below it (who we refer to as "the unsustainable"). We replicated this analysis in the context of policy support by

examining underestimations separately for supporters and detractors of the policy.

Pluralistic ignorance occurs when individuals significantly misperceive the prevalence of an opinion, norm, or behavior (Leviston, Walker and Morwinski 2013, Shamir and Shamir 1997, Prentice and Miller 1996). In extreme cases, people may mistakenly believe that most others endorse an opinion or behavior that, in fact, is supported only by a minority. Pluralistic ignorance can have behavioral consequences: individuals may act against their private personal norms in order to conform to a misperceived norm. For example, it has led climate-concerned individuals to self-silence (Geiger and Swim 2016). If underestimations are driven by pluralistic ignorance, then everyone, both the sustainable and the unsustainable, should underestimate others' sustainability. For instance, both supporters and detractors underestimated public support for offshore wind energy (Sokoloski, Markowitz and Bidwell 2018), and both climate change deniers and believers overestimated the proportion of deniers (Leviston, Walker and Morwinski 2013).

The false consensus effect, on the other hand, occurs when individuals overestimate the prevalence of their own opinions or behaviors (Ross, Greene and House 1977). Under this bias, individuals believe their views or actions are more common than they are perceived to be by those who hold different views or engage in different behaviors (Mullen *et al.* 1985). For example, both climate change believers and deniers perceived their own opinion on climate change as the most common (Leviston, Walker and Morwinski 2013). Similarly, climate experts tend to overestimate how widely their beliefs about climate outcomes are shared within the expert community (Wynes *et al.* 2024). If our findings are driven by the false consensus effect rather than pluralistic ignorance, then only the unsustainable participants should underestimate others' sustainability.

However, if both pluralistic ignorance and the false consensus effect are at play, then both sustainable and unsustainable participants should underestimate others' sustainable behaviors, personal norms, and policy support; and the unsustainable should do so to a greater extent. Indeed, prior research has found that although misperceptions about others' pro-environmental support are widespread, they tend to be especially pronounced among certain groups, such as climate change deniers (Leviston, Walker and Morwinski 2013), Republicans (Andre *et al.* 2024b, Sparkman, Geiger and Weber 2022), individuals living in regions with higher support for Trump, and those residing in areas with fewer environmental protests (Sparkman, Geiger and Weber 2022).

Our findings provide evidence of both pluralistic ignorance and the false consensus effect. The false consensus effect appears strong and widespread: underestimations of sustainable behaviors and personal norms were significantly larger among the unsustainable, and the share of underestimators was also higher in this group (on average, 76% and 70% of unsustainable individuals underestimated

others' behaviors and personal norms, respectively, compared to 40% and 48% among sustainable individuals). A similar pattern emerged for policy support: policy detractors underestimated the number of supporters significantly more than supporters did, and the proportion of underestimators was again higher among detractors (on average, 90% of detractors underestimated support, compared to 66% of supporters).

The evidence for pluralistic ignorance was more mixed, particularly regarding misperceptions of behaviors and personal norms. Sustainable participants underestimated others' vegetarian consumption and reduction of home heating but overestimated others' use of public transportation and the monetary amount donated. Underestimations of personal norms were driven exclusively by the unsustainable participants and were not significant among the sustainable. Pluralistic ignorance was more evident in the context of policy support: both detractors and supporters significantly underestimated public support for all policies except one. Taken together, these findings provide evidence of the false consensus effect influencing social expectations across all behavioral domains. We also find evidence of pluralistic ignorance, but only with respect to empirical expectations and expected policy support, and not consistently across all behavioral domains.

Our data also suggest that, consistent with past literature, participants exhibit conditional preferences in their sustainable behaviors and policy support. A one-standard-deviation increase in empirical expectations was associated with a 0.4 to 0.5 standard deviation increase in self-reported behaviors, while a one-standard-deviation increase in normative expectations was associated with a 0.2 to 0.4 standard deviation increase. Notably, a one-percentage-point increase in the expected amount donated by others from the €250 lottery was associated with a 0.91-percentage-point increase in the amount one donated. Similarly, a one-percentage-point increase in the expected proportion of policy supporters corresponded to an increase of between 0.7 and 1 percentage point in one's own likelihood of supporting the policy.

An interesting question we explored is whether people's personal norms are also conditional on social expectations, or whether they operate independently. Personal norms are distinct from social norms and serve as complementary predictors of behavior (Bašić and Verrina 2024, De Groot, Bondy and Schuitema 2021, Bertoldo and Castro 2016). Indeed, in our data, a one-standard-deviation increase in personal norms was associated with a 0.3 to 0.6 standard deviation increase in self-reported behaviors. Given their influence on behavior, if social expectations also shape personal norms, then the effect of social expectations on behavior could operate through two pathways: a direct effect and an indirect effect via the updating of personal norms.

There is evidence that empirical expectations influence the formation and updating of individuals'

personal moral judgments, an effect referred to as the "common is moral" heuristic (Deutchman *et al.* 2024, Lindström *et al.* 2018). Based on this, we hypothesize that sustainable personal norms, like behaviors, may also be conditional on social expectations. Indeed, a one-standard-deviation increase in empirical expectations was associated with a 0.4 standard deviation increase in personal norms, while a one-standard-deviation increase in normative expectations was associated with a 0.6 to 0.8 standard deviation increase in personal norms. Taken together, these results suggest that participants exhibit conditional preferences in their everyday sustainable behaviors, personal norms, and policy support. While empirical expectations were more strongly associated with behavior, normative expectations showed a stronger association with personal norms.

Luxembourgish residents and workers represent a particularly interesting sample for studying misperceptions related to sustainable social norms, given the country's high income per capita. Luxembourg ranks as the second-highest country in the world in terms of average annual salary (MIMCO Capital 2023). According to a recent Oxfam report, the richest 1% of humanity is responsible for more carbon emissions than the poorest 66% (Khalfan *et al.* 2023). Moreover, while high-income countries account for 40% of global consumption-based carbon emissions, the contribution from low-income countries is a mere 0.4%. Not only are high-income-per-capita countries more polluting, but they are also less willing to contribute financially to the fight against climate change (Andre *et al.* 2024a). In line with this trend, Luxembourg is the top emitter per capita in the European Union and ranks 20th globally (Crippa *et al.* 2023). Given this, studying the barriers to behavioral change in high-income, high-emission countries like Luxembourg is particularly important for reducing global emissions.

The remainder of the paper is organized as follows: Section 2.2 describes the survey design, the measures collected, and the sample. Section 2.3 presents an overview of the main results. Section 2.4 concludes.

2.2 Study Design

The survey was part of a larger online study conducted in collaboration with the Ministry of Economy in Luxembourg (Verheyden *et al.* 2024), launched in November 2022. We contacted 3,700 volunteers who had previously participated in LISER surveys and had consented to be recontacted for future research. The volunteers were adult Luxembourg residents and cross-border workers (individuals residing in neighboring countries who work in Luxembourg). Among them, 1,292 participants completed the survey. The study was pre-registered at aspredicted.org/WYP_W6F

and received IRB approval from LISER.

We translated the survey into four languages: English, French, German, and Portuguese. Participants were informed that their participation was voluntary and that they could withdraw from the survey at any time. We also assured them that their responses would remain anonymous. Additionally, we informed them that the survey included two attention-check questions and that failing both would prevent them from continuing and completing the survey. To encourage participation, respondents received a fixed reward of €10. They could also earn a bonus ranging from €10 to €30 based on their scores in questions eliciting their social expectations. Specifically, participants received €10 if they ranked among the top 400 scorers, €20 if they ranked among the top 100, and €30 if they ranked among the top 50. Finally, 10 participants were randomly selected through a lottery to receive an additional prize of €250.

All participants were informed during the initial survey that they would be recontacted to participate in two additional waves, conducted in April and July 2023. These follow-up studies serve as the basis for the third chapter of this dissertation.

2.2.1 Measures

After collecting a wide range of demographic data from participants, we gathered self-reported data on behaviors, personal norms, and support for policies related to the three most carbon-saving actions: consuming vegetarian meals, reducing home heating, and using public transportation instead of cars. Participants also completed a donation task, which served as our measure of incentivized behavior. Finally, we elicited participants' social expectations regarding all the aforementioned measures. See Appendix A2.5 for the specific survey instructions.

Behaviors. We asked participants to report their current levels of animal protein consumption ("Over the last 7 days, how many meals containing meat, fish or seafood did you eat?"; scale ranging from 0 to 21),[1] home heating ("What is the usual temperature of your dwelling when you are at home when it is less than 10 degrees outside?"; scale ranging from 15 to 30), and use of public transportation ("Compared to a 30-minute car trip, what is the maximal trip duration you would accept if you were to systematically use public transport instead?"; open-text box). We used a hypothetical question for public transportation because mobility patterns in Luxembourg are highly dependent on individual constraints (e.g., place of residence, workplace location, and access to reliable

[1]We further informed participants that by "one meal containing meat, fish or seafood", we mean a meal containing at least 50 grams of it.

public transport). This standardization is particularly important when eliciting personal norms and normative expectations. Therefore, we chose to present a scenario in which all participants faced the same trade-off. To ensure consistency in the directionality of our measures, we reverse-coded responses on animal protein consumption (21 - response) to obtain a measure of vegetarian meals per week. Similarly, we reverse-coded responses on home temperature (30 - response) to capture "degrees below the maximum." This approach ensures that higher values consistently reflect more sustainable consumption behaviors.

Additionally, we offered participants the opportunity to donate part of their potential €250 lottery prize to a carbon offset project with a slider ranging from 0 (labeled "all to myself") to 250 (labeled "all to carbon credits"). We informed them that if they won the lottery, the money would be allocated according to their decision. We also informed them that by donating the full amount, they might offset 4 months of carbon emissions.[2]

Personal norms. We asked participants to report, for each of the sustainable behaviors described above, what they considered to be the ethically appropriate level from a sustainability perspective. For example, regarding animal protein consumption, we asked: "Due to the amount of energy needed to produce them, meat, fish and seafood are not a sustainable source of proteins. With these sustainability concerns in mind, how many meals per week containing meat, fish or seafood do you think it would be ethically appropriate to eat?" (slider ranging from 0 to 21). We asked similar questions for the other two behaviors, using response scales identical to those used in the behavioral questions. For home temperature, we asked participants to indicate the maximum temperature they found ethically appropriate for well-insulated and poorly insulated dwellings separately, and then we averaged the answers to these two questions. Same as before, we reverse-coded responses on animal protein consumption and home temperature. As pre-registered, we did not ask this question for the donation decision.[3]

Empirical and normative expectations. We elicited empirical and normative expectations by asking participants to estimate the responses given by all other participants in the survey, following a method similar to that used by Bicchieri *et al.* (2022) and Krupka and Weber (2013). Specifically, for each question, we asked: "Please guess what will be the most frequent answer in the survey

[2]Prior to the lottery donation decision, participants were also given the opportunity to donate their survey earnings (fixed payment and bonus) to a carbon offset project (results not reported in this paper). Thus, it is important to note that the lottery donation task was the second donation decision presented in the survey.

[3]The donation decision was presented at the very end of the survey, after thanking participants for their previous answers. Due to the length of the survey and potential participant fatigue, we did not want to ask too many questions in this section.

to the question...". Estimates of behaviors served as our measure of empirical expectations, while estimates of personal norms constituted our measure of normative expectations. These questions were incentivized: participants could earn points based on the accuracy of their estimates, which in turn determined their chances of receiving a bonus payment. The only exception was the estimation of others' donation amount, which was not incentivized.[4] At the beginning of the survey, all participants were informed that we aimed to recruit a maximum of 1,500 participants residing in Luxembourg or its neighboring areas.

Policy support. Participants indicated whether they would support six hypothetical policies aimed at promoting the sustainable behaviors by restricting the associated unsustainable ones. We presented two types of policies for each behavioral domain: a regulation and a tax. For vegetarian consumption, we proposed a reduction in the availability of red meat and an increase in VAT on meat, fish, and seafood to 17%. For home heating, we presented a rationing of fossil energy sources and a 10% tax on rental income received by landlords who rent out accommodations with insufficient energy efficiency. For public transportation, we proposed a ban on cars in densely populated areas and a €5 toll on Luxembourg's highways. All policies were described as being accompanied by measures to promote the more sustainable alternatives.

Expected policy support. We asked participants to estimate, out of 100, how many other participants supported each of the six policies. As before, these questions were incentivized, and participants could earn points depending on how accurate their estimates were, which determined their chances of receiving a bonus payment.

2.2.2 Concerns for social desirability bias, experimenter demand effects and self-selection

We followed a protocol similar to that of Andre *et al.* (2024b) to address concerns about social desirability bias. Participants were informed that their responses were anonymous, and their beliefs about others were incentivized. Therefore, even if their self-reported behaviors were inflated due to social desirability, they could anticipate this bias when estimating the responses of others, thereby mitigating concerns about the validity of any detected misperceptions. Furthermore, a large-scale meta-analysis by Kormos and Gifford (2014) concluded that self-reported pro-environmental behaviors

[4]This is because, prior to the lottery donation decision, participants had already made a donation decision involving their bonus payments, and all bonus-related earning opportunities needed to be presented in advance.

are strongly associated with objective measures. Finally, Reisinger (2022) provided evidence that anonymity in self-reports greatly reduce social desirability bias.

According to De Quidt, Haushofer and Roth (2018), participants making incentivized and non-incentivized choices in online studies responded similarly to experimenter demand. Likewise, non-representative online samples were no more susceptible to experimenter demand than representative ones. The authors concluded that, at least in anonymized online surveys, concerns about experimenter demand effects are limited.

To mitigate concerns about self-selection bias and the potential overrepresentation of participants with strong pro-environmental preferences, we obscured the true purpose of the survey when inviting participants. Specifically, we informed them only that the survey would address their perceptions and behaviors related to societal challenges facing the Grand Duchy of Luxembourg. In addition, strong monetary incentives to participate help mitigate concerns that only individuals with strong social preferences would take part. Finally, existing evidence suggests that the use of volunteer samples has a negligible impact on participants' social preferences and behavioral patterns (see, for example, Anderson *et al.* (2013); Falk, Meier and Zehnder (2013); Abeler and Nosenzo (2015)).

2.2.3 Hypotheses

We pre-registered the following main hypotheses:

H1. Inaccurate empirical expectations: Participants will underestimate others' sustainable behaviors and monetary donations.

H2. Inaccurate normative expectations: Participants will underestimate others' personal norms concerning sustainable behaviors.

H3. Inaccurate expected policy support: Participants will underestimate others' support for policies that restrict unsustainable behaviors.

Additionally, we were also interested in examining whether individuals may have conditional preferences in their behaviors, personal norms and policy support. Hence, we pre-registered the following secondary hypotheses:

H4. Personal norms, empirical expectations and normative expectations positively predict sustainable behaviors and monetary donations.

H5. Empirical and normative expectations positively predict personal norms concerning sustainable behaviors.

H6. Expected policy support positively predicts one's own policy support.

2.2.4 Sample

Table 3.1 presents the demographic statistics of the 1,292 respondents who participated in the survey. Income information was collected using intervals of €2,000. The median total net household monthly income falls within the €6,000-€8,000 interval. Based on this, we define "low income" as less than €6,000 and "high income" as more than €8,000. Since our sample is not representative of the Luxembourgish population, we applied weights to adjust for the population's age structure and gender composition in Table 1 and in all our subsequent analyses.

Table 2.1: Demographic distribution (weighted)

	Summary (N=1,292)
Low income (< 6,000€)	0.450 (0.498)
High income (> 8,000€)	0.279 (0.449)
Aged below 35	0.326 (0.469)
Aged above 65	0.156 (0.363)
Female	0.494 (0.500)
Higher education	0.653 (0.476)
Living in urban area	0.482 (0.500)

Mean proportion (SD in parentheses).

2.3 Results

As described earlier, all our analyses were conducted using weights to adjust our sample to be representative of Luxembourg in terms of age and gender.[5] In all analyses, we included income, age, gender, and education as control variables. Appendix A2.1 provides an overview of which analyses reported in this section are pre-registered. Appendices A2.2 and A2.3 present replications of our pre-registered analyses without weights and without controls as a robustness check. As shown in Appendix A2.2, weighting the data has minimal impact on our measures of interest and results, with mean responses differing by only a few decimal places.

[5]We did not consider additional demographic variables when creating our weights for two reasons: 1) data were not always readily available, and 2) each additional variable would inflate the variance of our weights and estimates, and we wanted to avoid excessive increases in variance.

A post-hoc power analysis determined that, with our sample size, we had over 80% power to detect effect sizes as small as Cohen's d = 0.08 with $\alpha = 0.05$ in our analyses (Cohen 1988).[6]

2.3.1 Underestimation of others' sustainable behaviors, personal norms, and policy support

Result 1a. Participants significantly underestimate others' sustainable behaviors and personal norms.

Table 2.2: Underestimation of sustainable behaviors and norms (weighted)

	Own's answer (mean)	Expectation others (mean)	Cohen's d
Behaviors			
Vegetarian meals	14.69 meals	12.82 meals	0.50
Lower home temperature	9.88°	9.04°	0.57
Public mobility time	45.79 min	42.88 min	0.17
Donation of lottery gains	31.68%	28.16%	0.12
Personal Norms			
Vegetarian meals	16.52 meals	15.75 meals	0.28
Lower home temperature	9.61°	9.31°	0.20
Public mobility time	57.45 min	55.21 min	0.13

As shown in Table 2.2, our respondents misperceived what other survey participants did and considered appropriate to do, underestimating others' sustainability levels both in behavior and in personal norms. This pattern was consistent across all behavioral domains. The effect sizes of these underestimations varied considerably, ranging from small (Cohen's d = 0.1) to medium-large (Cohen's d = 0.6; Cohen 1988). On average, underestimations of behaviors had an effect size of Cohen's d = 0.34 and underestimations of personal norms had an effect size of Cohen's d = 0.20.

To assess whether the underestimation of others' sustainable behaviors and personal norms was statistically significant, we created an underestimation measure by calculating the gap between participants' social expectations and the actual mean responses for behaviors and personal norms across all participants. We then estimated the average underestimations, controlling for age, gender, income, and education. As shown in Figure 2.1, all underestimations are statistically significant at the 95% confidence level. On average, participants underestimated the number of vegetarian meals

[6]According to standard conventions, Cohen's d = 0.2 is defined as a small effect size (Cohen 1988).

others consumed per week by 1.56 meals ($p < 0.001$) and the number they deemed appropriate by 0.64 meals ($p < 0.001$). They underestimated the reduction in home temperature by 0.89°C ($p < 0.001$) and the appropriate reduction by 0.41°C ($p < 0.001$). They also underestimated the number of minutes others were willing to sacrifice to switch from car to public transport by 1.94 minutes ($p < 0.001$), and the number of minutes deemed appropriate by 1.55 minutes ($p = 0.013$). Finally, they underestimated the average share of the lottery donation by 2 percentage points ($p = 0.001$).

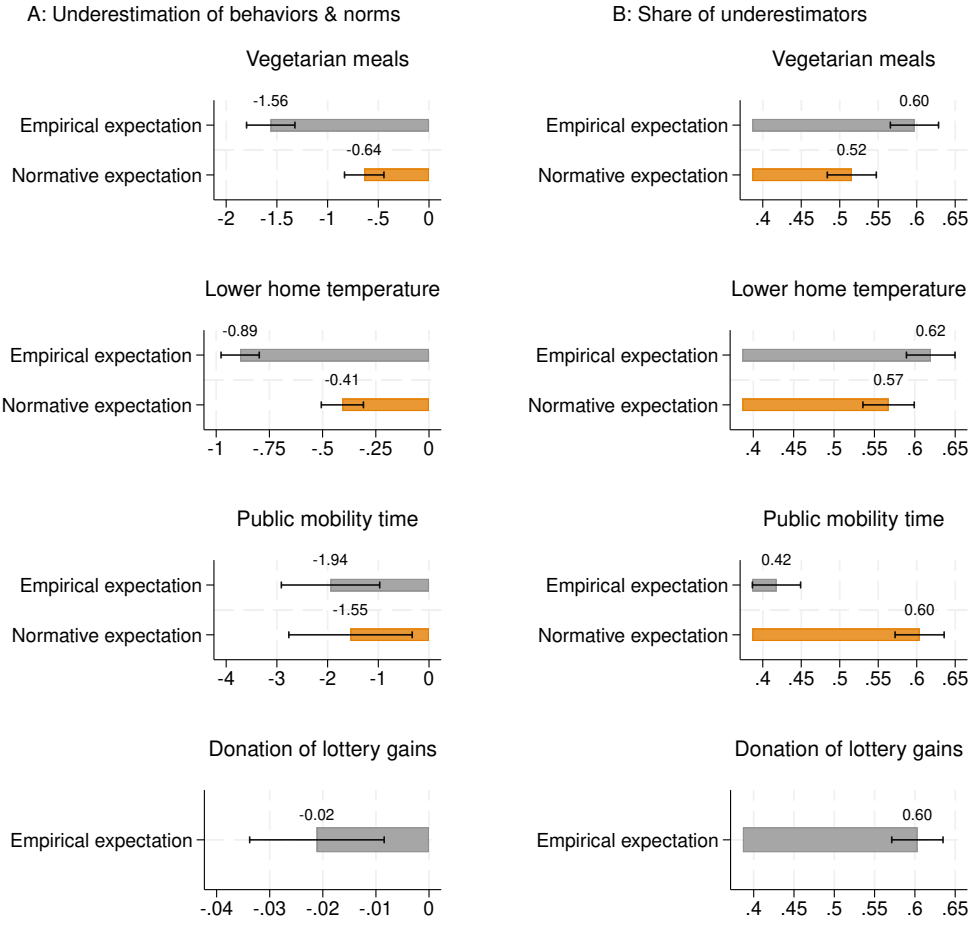


Figure 2.1: **Panel A** displays the mean underestimation of others' sustainable behaviors and personal norms (expectation minus mean response). **Panel B** shows the share of underestimators for each sustainable behavior and personal norm. In both panels, estimates control for age, gender, income, and education. Confidence intervals are shown at the 95% level.

We also examined the share of underestimators for each of the measures above, controlling for gender, age, income, and education. As shown in Figure 2.1, the majority of participants were pessimistic in their predictions. Specifically, 60% and 52% of participants underestimated the

number of vegetarian meals others consumed and the number they deemed appropriate to consume, respectively. Similarly, 62% and 57% underestimated others' reduction in home temperature and the appropriate reduction. Notably, only 42% of participants underestimated the minutes others were willing to sacrifice to use public transport, but 60% underestimated the minutes others deemed appropriate to sacrifice. Finally, 60% underestimated the average share of lottery donations. On average, 56% of participants underestimated others' behaviors, and 56% also underestimated others' personal norms.

Result 1b. Participants significantly underestimate others' policy support.

Table 2.3: Underestimation of support for restrictive and taxing policies (weighted)

	Supporters (%)	Expected supporters (%)	Cohen's d
Regulation on red meat	63.50	35.12	0.78
VAT on meat	36.89	24.17	0.35
Fossil fuel rationing	52.53	35.93	0.43
Rental tax on poor insulation	63.86	43.71	0.53
Ban on cars in city center	50.19	33.67	0.43
Toll on highways	21.35	19.16	0.07

Misperceptions also emerged in participants' support for restrictive policies: they underestimated the number of supporters for all six policies. These underestimations exhibited larger effect sizes than those observed for behaviors and norms, ranging from small (Cohen's d = 0.1) to large (Cohen's d = 0.8). On average, underestimations of policy support had an effect size of Cohen's d = 0.43. Additionally, participants expected that none of the six policies would be supported by a majority; in reality, four of them were.

To assess whether the underestimation of others' policy support is statistically significant, we created a measure of underestimation by calculating the gap between participants' expected share of supporters and the actual share of supporters. We then estimated the average underestimations, controlling for the same variables as described earlier. As displayed in Figure 2.2, all underestimations are statistically significant at the 95% confidence level. On average, participants underestimated support for a regulation on red meat by 26 percentage points ($p < 0.001$), for an increase in VAT on meat, fish, and seafood by 13 percentage points ($p < 0.001$), for a rationing policy on fossil fuels by 15 percentage points ($p < 0.001$), for a rental tax on poorly insulated homes by 21 percentage points ($p < 0.001$), for a ban on cars in city centers by 17 percentage points ($p < 0.001$), and for a

highway toll by 3 percentage points ($p < 0.001$).

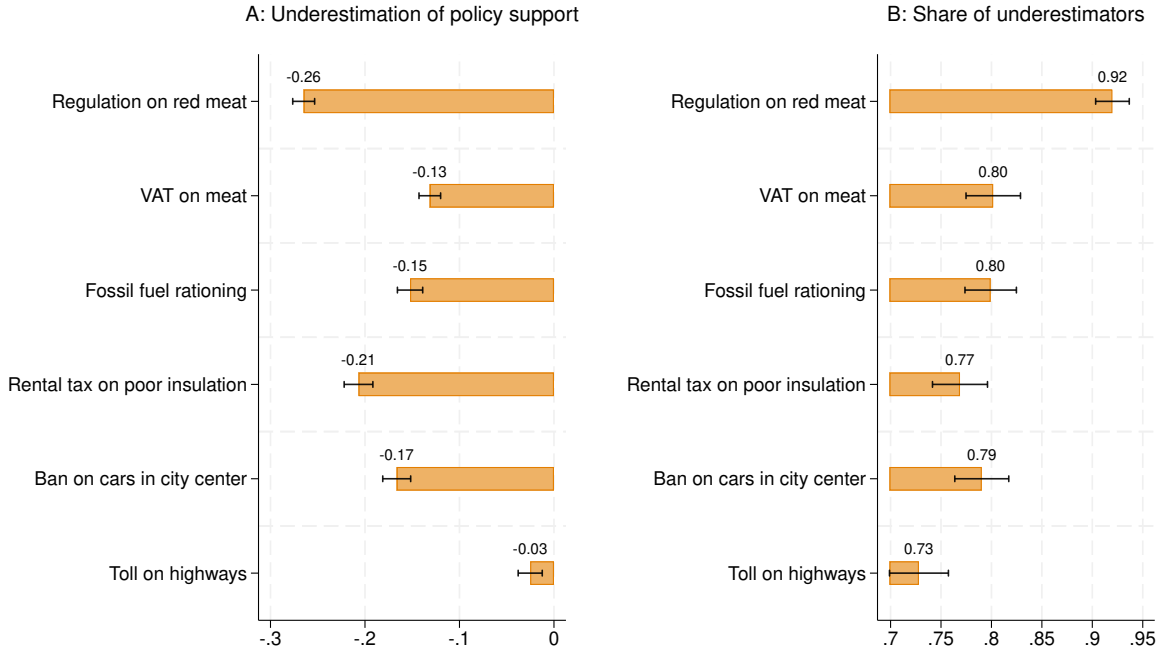


Figure 2.2: **Panel A** displays the mean underestimation of other's policy support (expected % of supporters - actual % of supporters). **Panel B** shows the share of underestimators for each policy. In both panels, estimates control for age, gender, income, and education. Confidence intervals are reported at the 95% level.

As shown in Figure 2.2, and controlling for the same variables described earlier, the share of underestimators is also significantly higher for policy support than for behaviors and personal norms. Specifically, 93% of participants underestimated support for a regulation on red meat; 80% for an increase in VAT on meat, fish, and seafood; 80% for a rationing policy on fossil fuels; 77% for a rental tax on poorly insulated homes; 79% for a ban on cars in city centers; and 73% for a highway toll. On average, 80% of participants underestimated others' policy support.

2.3.2 Heterogeneity analyses: examining the false consensus effect and pluralistic ignorance

Result 2a. The underestimation of sustainable behaviors is consistent with both the false consensus effect and pluralistic ignorance, whereas the underestimation of personal norms is consistent only with the false consensus effect.

We created a dummy variable that takes the value 1 if an individual reported a sustainable behavior level above the average (referred to as a "sustainable" individual), and 0 if they reported a level below the average (referred to as an "unsustainable" individual). Note that we use the terms sustainable and unsustainable only in relation to the specific behavior in question. Thus, the same participant could be classified as sustainable in vegetarian consumption but unsustainable in home heating. We then replicated all previous analyses including this dummy variable in the model (regressions reported in Appendix A2.4).

As shown in Figure 2.3, the empirical expectations of sustainable participants were mixed. They were pessimistic about others' engagement in vegetarian meal consumption ($p = 0.004$) and home temperature reduction ($p < 0.001$), but optimistic regarding public mobility time ($p = 0.001$) and lottery donations ($p < 0.001$). However, even when they underestimated others, the degree of underestimation was significantly lower than that of unsustainable participants (see Columns 1 and 3 in Table A2.7 in Appendix A2.4, all $p < 0.001$). Moreover, with the exception of home heating reduction, the share of underestimators among sustainable individuals did not constitute a majority for any behavior. In contrast, a large majority of unsustainable individuals were underestimators. On average, 40% of sustainable individuals underestimated others' behaviors, compared to 76% of unsustainable individuals. Logistic regression analyses confirmed that the share of underestimators was significantly higher among unsustainable individuals than among sustainable ones for all behaviors (all $p < 0.001$).

When it comes to personal norms, underestimations were exclusively driven by unsustainable individuals. The misperceptions of sustainable participants were not significantly different from zero (all $p > 0.05$). Again, with the exception of public mobility time, the share of underestimators among sustainable individuals did not constitute a majority for any personal norm, whereas a majority of unsustainable individuals underestimated others' personal norms. On average, 48% of sustainable individuals underestimated others' personal norms, compared to 70% of unsustainable individuals. Logistic regression analyses confirmed that the share of underestimators was significantly higher among unsustainable individuals than among sustainable ones for all personal norms (all $p < 0.001$).

Result 2b. The underestimation of policy support is consistent with both the false consensus effect and pluralistic ignorance.

We replicated the same heterogeneity analyses for policy support, distinguishing between individuals who supported a policy (referred to as supporters) and those who did not (referred to as detractors), with regressions reported in Appendix A2.4. As before, the same individual could be

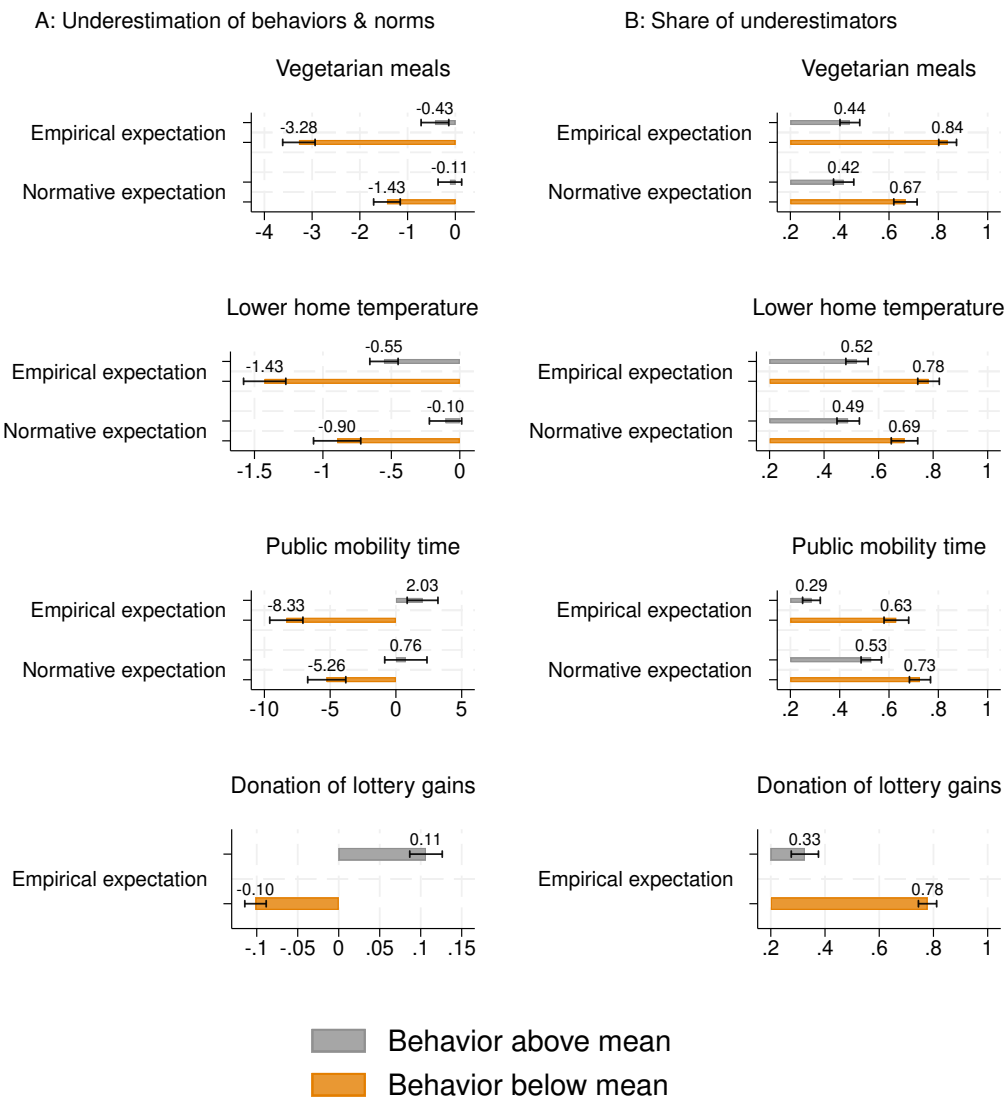


Figure 2.3: **Panel A** displays the mean underestimation of others' sustainable behaviors and personal norms (expectation minus mean response). **Panel B** shows the share of underestimators for each sustainable behavior and personal norm. In both panels, estimates control for age, gender, income, and education, and are separated by individuals whose behavior levels are above or below the mean. Confidence intervals are shown at the 95% level.

labeled a supporter of one policy and a detractor of another.

With the exception of the highway toll, both supporters and detractors significantly underestimated the share of supporters (all $p < 0.001$). However, detractors underestimated support to a greater extent than supporters (see Columns 1–5 in Table A2.8 in Appendix A2.4, all $p < 0.001$). For the highway toll policy, both groups significantly overestimated their own prevalence (both $p < 0.001$).

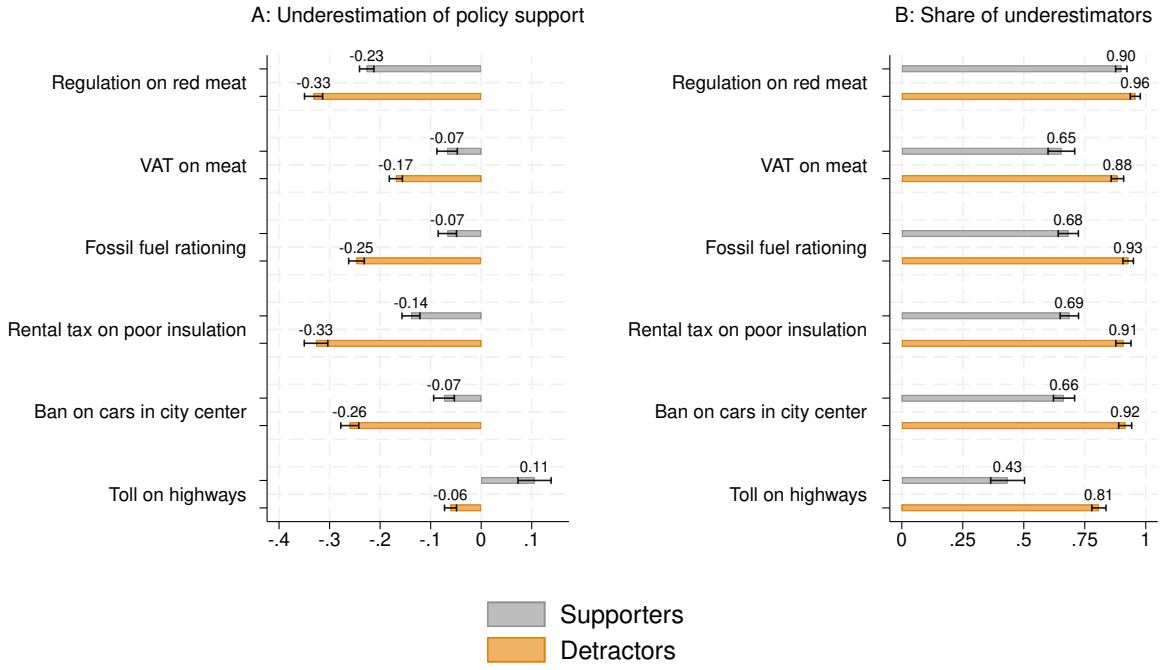


Figure 2.4: **Panel A** displays the mean underestimation of other's policy support (expected % of supporters - actual % of supporters). **Panel B** shows the share of underestimators for each policy. In both panels, estimates control for age, gender, income, and education, and are separated by supporters or detractors of a policy. Confidence intervals are reported at the 95% level.

Again, with the exception of the highway toll, a majority of supporters underestimated others' support across all policies. On average, 66% of supporters underestimated support, compared to 90% of detractors. Logistic regression analyses confirmed that the share of underestimators was significantly higher among policy detractors than supporters for all policies (all $p < 0.01$).

2.3.3 Conditional preferences in behaviors, personal norms, and policy support

As pre-registered, we assessed whether individuals exhibit conditional preferences in their sustainable behaviors and policy support by examining whether social expectations are positively associated with these outcomes. We also evaluated whether personal norms serve as significant predictors. Since empirical and normative expectations are conceptually related, we followed the approach of Andre *et al.* (2024b) and estimated separate models for each predictor. For ease of interpretation, all behavior and personal norm variables were standardized to have a mean of zero and

a standard deviation of one. Policy support and lottery donations, however, were not standardized, as they are expressed in percentage points, which are already easily interpretable.

Result 3. Social expectations and personal norms positively predict behaviors and policy support.

Tables 2.4 and A2.5 show the results of our regression analyses. As expected, both social expectations and personal norms positively predicted one's behavior (all $p < 0.001$).

Table 2.4: Do social expectations predict behaviors and personal norms?

	Vegetarian meals		Lower home temperature		Public mobility time		Donation of lottery gains
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Behaviors							
Empirical expectation	0.421***		0.405***		0.493***		0.906***
Normative expectation		0.261***		0.363***		0.237***	
Panel B: Behaviors							
Personal norm	0.570***		0.567***		0.289***		
Panel C: Personal norms							
Empirical expectation	0.362***		0.389***		0.419***		
Normative expectation		0.560***		0.692***		0.796***	
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1292	1292	1292	1292	1292	1292	1292

* p<0.05, ** p<0.01, *** p<0.001.

Results from OLS regressions with weighted data. Coefficients in Columns 1-6 are expressed as SD. Coefficient in Column 7 is expressed as percentage points. Panel A includes empirical expectations and normative expectations as predictors of behaviors. Panel B includes personal norms as predictors of behaviors. Panel C includes empirical expectations and normative expectations as predictors of personal norms.

As shown in Panel A of Table 2.4, a 1-standard-deviation increase in empirical expectations is associated with an increase in self-reported behaviors of 0.4 to 0.5 standard deviations, and a 1-standard-deviation increase in normative expectations is associated with an increase in self-reported behaviors of 0.2 to 0.4 standard deviations (Panel A, Columns 1-6). We also found that a 1-percentage-point increase in the expected amount donated by others of the €250 lottery is associated with a 0.91-percentage-point increase in the amount one donated (Panel A, Column 7). Finally, we also found that personal norms are significantly associated with daily sustainable behaviors. A 1-standard-deviation increase in personal norms is associated with an increase in behaviors of 0.3 to 0.6 standard deviations.

Additionally, as shown in Table A2.5, a 1-percentage-point-increase in the expected proportion of supporters of a policy corresponded to a minimum of 0.7-percentage-point increase and a maximum of 1-percentage-point increase in one's own probability to support a policy.

Table 2.5: Does expected policy support predict one's own support for policies?

	Regulation on red meat (1)	VAT on meat (2)	Fossil fuel rationing (3)	Rental tax on poor insulation (4)	Ban on cars in city center (5)	Toll on highways (6)
Expected support	0.709***	0.786***	1.04***	0.775***	0.951***	0.855***
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1292	1292	1292	1292	1292	1292

* p<0.05, ** p<0.01, *** p<0.001.

Results from OLS regressions with weighted data. Coefficients expressed as percentage points.

Result 4. Social expectations positively predict personal norms.

Following our pre-registration, we also assessed whether people may have conditional personal norms by examining whether social expectations are positively associated with them. As shown in Panel C of Table 2.4, a 1-standard-deviation increase in empirical expectations is associated with a 0.4-standard-deviation increase in personal norms, and a 1-standard-deviation increase in normative expectations is associated with an increase in personal norms of 0.6 to 0.8 standard deviations (all $p < 0.001$).

2.3.4 Discussion

We confirmed our pre-registered hypotheses by finding that participants significantly underestimated others' everyday sustainable behaviors, personal norms, and support for policies targeting such behaviors, with effect sizes ranging from small to large. We also replicated the findings of Andre *et al.* (2024b), observing that participants significantly underestimated how much others donate to a carbon-offset project. Notably, the magnitude of this underestimation was much smaller in our study compared to theirs. While they reported a 40-percentage-point gap between expected and actual donation amounts, we observed only a 2-percentage-point gap.[7] Similarly, whereas 86% of participants in their study underestimated others' donations, the corresponding share in our sample was 60%. This difference may be explained by the fact that participants in our study were first given the opportunity to donate their survey earnings, making this the second donation task they encountered and potentially reducing the total donation amount. However, participants may also have anticipated this when forming expectations about others. In any case, the underestimation remained significant.

[7]Notably, the lottery size offered by Andre *et al.* (2024b) was also larger: $450, compared to €250 in our study.

Although our participants significantly underestimated others' behaviors and personal norms, the size of the underestimation was larger for behaviors. This is evident in both the effect sizes reported in Table 2.2 and in Figure 2.1. This contrasts with the findings of Andre *et al.* (2024b), who observed that personal norms were underestimated to a greater extent. This result is especially interesting given that behaviors should be more observable than norms. Additionally, while underestimations were significant across all three main behavioral domains, they were greatest for vegetarian consumption and reduction in home heating.

The share of participants in our sample who underestimated others' everyday sustainable behaviors and personal norms (on average, 56% for both) is smaller than the proportion found by Andre *et al.* (2024b), who reported that 67% of participants underestimated how many others try to fight climate change and 76% underestimated how many others believe one should try to fight climate change. This is expected, as publicly observable behaviors, especially those with objective measures at the intensive level, are harder to misperceive.

Underestimations were larger for support of restrictive policies targeting behavior than for behaviors and personal norms, and, once again, they were greatest for vegetarian consumption and reduction in home heating. On average, 80% of participants underestimated others' policy support. The mean effect size for underestimations of behaviors and personal norms was Cohen's d = 0.34 and Cohen's d = 0.20, respectively, whereas the mean effect size for underestimations of policy support was Cohen's d = 0.43. Notably, participants underestimated support for regulations on red meat by nearly half. Overall, the size of policy support underestimations was smaller than the size found by Sparkman, Geiger and Weber (2022) regarding underestimation of support for climate policies. This is not surprising, as Sparkman et al.'s policies were primarily costly for companies or governments, with minimal personal cost for individuals. As predicted, our policies, which involve a degree of personal sacrifice for individuals, show a lower level of support and a lower estimated level of support. However, the share of participants who underestimated the prevalence of support is similar.

Examining the average underestimations of sustainable and unsustainable participants separately allows us to assess the presence of pluralistic ignorance and the false consensus effect. If the most sustainable participants in our sample underestimated others' behaviors and personal norms, the results point to pluralistic ignorance. If, instead, they overestimated others, the results are driven by the false consensus effect. If their misperceptions were not significantly different from zero, the results again indicate a false consensus effect, but only among the unsustainable. Finally, if the sustainable participants also underestimated others, but to a lesser extent than the unsustainable, the findings reflect a combination of pluralistic ignorance and the false consensus effect.

For empirical expectations, we find clear evidence of the false consensus effect across all behavioral domains, especially in public mobility time and lottery donations, where both groups expected others to behave like themselves. For vegetarian consumption and home heating reduction, we again find evidence of the false consensus effect, but only among the unsustainable, alongside signs of pluralistic ignorance. That is, both sustainable and unsustainable individuals significantly underestimated others' vegetarian consumption and home heating reduction, but the unsustainable did so to a greater extent. Notably, on average, 76% of unsustainable individuals underestimated others' behaviors, compared to 40% of sustainable individuals (this share increases slightly to 48% when considering only vegetarian consumption and home heating reduction).

For normative expectations, underestimations were driven exclusively by the unsustainable, while misperceptions among the sustainable were not significantly different from zero. Hence, we observe the false consensus effect only among the unsustainable, with no evidence of pluralistic ignorance. On average, 70% of unsustainable individuals underestimated others' personal norms, whereas 48% of sustainable individuals did so.

When examining the misperceptions of policy supporters and detractors separately, we find much clearer evidence of pluralistic ignorance: both groups underestimated others' support for all policies except one. Similarly, a majority of both supporters and detractors were underestimators, on average, 66% and 90%, respectively. We also observe evidence of the false consensus effect among detractors, who underestimated policy support to a greater extent than supporters. Evidence of the false consensus effect from both groups appears only in the case of the highway toll policy, where both supporters and detractors overestimated their own prevalence.

These results suggest that while the underestimation of behaviors and policy support seem to be driven by both the false consensus effect and pluralistic ignorance, the underestimation of personal norms seems to be driven solely by the false consensus effect.

Finally, our results suggest that individuals may hold conditional preferences regarding their sustainable behaviors, personal norms, and policy support, consistent with past research. Notably, Andre *et al.* (2024a) found that a 1-percentage-point increase in the perceived proportion of others willingness to donate was associated with a 0.46-percentage-point increase in one's own likelihood of contributing. Building on this, we found that a 1-percentage-point increase in the expected amount donated by others corresponded to a 0.91-percentage-point increase in the amount donated by the participant.

Similarly, a 1-percentage-point increase in the expected proportion of policy supporters was associated with an increase of between 0.7 and 1 percentage point in an individual's own likelihood

of supporting the policy. These effect sizes are substantially larger than those reported by Andre *et al.* (2024a) regarding donation behavior.

Our findings also suggest that social expectations may influence behavior both directly and indirectly, by shaping personal norms. While empirical expectations were stronger predictors of behavior, normative expectations more strongly predicted personal norms.

2.4 Conclusion

In this paper, we document systematic underestimations among Luxembourgers regarding the extent to which others engage in everyday sustainable behaviors, the level of engagement they deem appropriate, and their support for policies restricting and taxing unsustainable behaviors. We focus on the three domains with the highest potential for carbon emission reduction: vegetarian meal consumption, home heating reduction, and the use of public transportation (Hitaj, Igos and Gibon 2022).

Our findings provide evidence that both pluralistic ignorance and the false consensus effect contribute to these underestimations. Specifically, both mechanisms seem to drive the underestimation of behaviors and policy support, whereas the underestimation of personal norms appears to be driven solely by the false consensus effect. Consistent with this, the largest underestimations were found among the most unsustainable individuals and the detractors of the policies.

Finally, in line with past research, our results suggest that individuals hold conditional preferences regarding their everyday sustainable behaviors and policy support. They also suggest that social expectations may influence behavior through both a direct channel and an indirect one, by shaping personal norms.

Together, these findings indicate that correcting misperceptions about others' behaviors, personal norms, and policy support may be a promising avenue for promoting sustainability, consistent with previous evidence (Andre *et al.* 2024b, Geiger and Swim 2016). Policy interventions and communication strategies that make social norms and public support more visible could help shift individual behaviors toward more climate-friendly practices. In particular, information about social norms could boost engagement not only at the extensive margin (encouraging more people to adopt sustainable behaviors) but also at the intensive margin (motivating individuals to increase their efforts). This is especially important given that household behavior accounts for a significant share of global emissions, particularly in high-income countries (Dubois *et al.* 2019). Notably, underestimations were especially large for support of policies promoting behavioral change, making

informational campaigns on this topic a particularly promising policy target for advancing a sustainable transition to a low-carbon society.

Our results also underscore the importance of distinguishing between different types of social expectations, namely, empirical expectations (beliefs about what others do) and normative expectations (beliefs about what others deem appropriate to do). We find that the mechanisms driving inaccuracies in social expectations may differ between these two types. Therefore, a comprehensive study of misperceptions and barriers to behavioral change must consider both constructs and their interplay, not only with behaviors but also with important influencers, such as personal norms. Likewise, any belief-correction intervention should address both empirical and normative expectations (Bicchieri 2016).

A limitation of our current study is that, based on our survey data, it is difficult to disentangle conditional preferences from the false consensus effect. Are people conditional in their behavior and personal norms (influenced by their social expectations), or do they believe others behave and think like they do? According to past studies where social expectations were exogenously manipulated (Schultz, Khazian and Zaleski 2008, Nolan *et al.* 2008, Schultz 1999, Ferraro, Miranda and Price 2011, De Groot, Abrahamse and Jones 2013, Goldstein, Griskevicius and Cialdini 2007, Reese, Loew and Steffgen 2014, Kormos, Gifford and Brown 2015), participants have conditional preferences for sustainable behaviors. Therefore, our results align with existing literature. However, to establish a causal relationship between social expectations and behavior or personal norms, an experimental design would be required, something we address in the third chapter of this dissertation. Nonetheless, the results presented in this paper are promising for future work aimed at correcting misperceptions.

References

Abeler, Johannes, and Daniele Nosenzo. 2015. "Self-selection into laboratory experiments: pro-social motives versus monetary incentives." Experimental Economics, 18: 195–214.

Anderson, Jon, Stephen V Burks, Jeffrey Carpenter, Lorenz Götte, Karsten Maurer, Daniele Nosenzo, Ruth Potter, Kim Rocha, and Aldo Rustichini. 2013. "Self-selection and variations in the laboratory measurement of other-regarding preferences across subject pools: evidence from one college student and two adult samples." Experimental Economics, 16: 170–189.

Andre, Peter, Teodora Boneva, Felix Chopra, and Armin Falk. 2024a. "Globally representative evidence on the actual and perceived support for climate action." Nature Climate Change, 14(3): 253–259.

Andre, Peter, Teodora Boneva, Felix Chopra, and Armin Falk. 2024b. "Misperceived social norms and willingness to act against climate change." Review of Economics and Statistics, 1–46.

Bašić, Zvonimir, and Eugenio Verrina. 2024. "Personal norms—and not only social norms—shape economic behavior." Journal of Public Economics, 239: 105255.

Bertoldo, Raquel, and Paula Castro. 2016. "The outer influence inside us: Exploring the relation between social and personal norms." Resources, Conservation and Recycling, 112: 45–53.

Bicchieri, Cristina. 2005. The grammar of society: The nature and dynamics of social norms. Cambridge University Press.

Bicchieri, Cristina. 2016. Norms in the wild: How to diagnose, measure, and change social norms. Oxford University Press.

Bicchieri, Cristina, Eugen Dimant, Simon Gächter, and Daniele Nosenzo. 2022. "Social proximity and the erosion of norm compliance." Games and Economic Behavior, 132: 59–72.

Cialdini, Robert B, and Ryan P Jacobson. 2021. "Influences of social norms on climate change-related behaviors." Current Opinion in Behavioral Sciences, 42: 1–8.

Clayton, Susan, Patrick Devine-Wright, Paul C Stern, Lorraine Whitmarsh, Amanda Carrico, Linda Steg, Janet Swim, and Mirilia Bonnes. 2015. "Psychological research and global climate change." Nature climate change, 5(7): 640–646.

Cohen, Jacob. 1988. Statistical power analysis for the behavioral sciences. routledge.

Crippa, Monica, Diego Guizzardi, Federico Pagani, Manjola Banja, Marilena Muntean, Edwin Schaaf, W Becker, Fabio Monforti-Ferrario, Roberta Quadrelli, A Risquez Martin, *et al.* 2023. "GHG emissions of all world countries." Publications Office of the European Union, Luxembourg, 10: 953322.

Dana, Jason, Roberto A Weber, and Jason Xi Kuang. 2007. "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness." Economic Theory, 33(1): 67–80.

De Groot, Judith IM, Krista Bondy, and Geertje Schuitema. 2021. "Listen to others or yourself? The role of personal norms on the effectiveness of social norm interventions to change pro-environmental behavior." Journal of Environmental Psychology, 78: 101688.

De Groot, Judith IM, Wokje Abrahamse, and Kayleigh Jones. 2013. "Persuasive normative messages: The influence of injunctive and personal norms on using free plastic bags." Sustainability, 5(5): 1829–1844.

De Quidt, Jonathan, Johannes Haushofer, and Christopher Roth. 2018. "Measuring and bounding experimenter demand." American Economic Review, 108(11): 3266–3302.

Deutchman, Paul, Gordon Kraft-Todd, Liane Young, and Katherine McAuliffe. 2024. "People update their injunctive norm and moral beliefs after receiving descriptive norm information." Journal of Personality and Social Psychology.

Druckman, Angela, and Tim Jackson. 2016. "Understanding households as drivers of carbon emissions." Taking stock of industrial ecology, 181–203.

Dubois, Ghislain, Benjamin Sovacool, Carlo Aall, Maria Nilsson, Carine Barbier, Alina Herrmann, Sébastien Bruyère, Camilla Andersson, Bore Skold, Franck Nadaud, *et al.* 2019. "It starts at home? Climate policies targeting household consumption and behavioral decisions are key to low-carbon futures." Energy Research & Social Science, 52: 144–158.

Falk, Armin, Stephan Meier, and Christian Zehnder. 2013. "Do lab experiments misrepresent social preferences? The case of self-selected student samples." Journal of the European Economic Association, 11(4): 839–852.

Ferraro, Paul J, Juan Jose Miranda, and Michael K Price. 2011. "The persistence of treatment effects with norm-based policy instruments: evidence from a randomized environmental policy experiment." American Economic Review, 101(3): 318–322.

Geiger, Nathaniel, and Janet K Swim. 2016. "Climate of silence: Pluralistic ignorance as a barrier to climate change discussion." Journal of Environmental Psychology, 47: 79–90.

Goldstein, Noah J, Vladas Griskevicius, and Robert B Cialdini. 2007. "Invoking social norms: A social psychology perspective on improving hotels' linen-reuse programs." Cornell Hotel and Restaurant Administration Quarterly, 48(2): 145–150.

Hertwich, Edgar G, and Glen P Peters. 2009. "Carbon footprint of nations: a global, trade-linked analysis." Environmental science & technology, 43(16): 6414–6420.

Hitaj, Claudia, Elorri Igos, and Thomas Gibon. 2022. "Towards decabornisation: Understanding and reducing our carbon footprint in Luxembourg." Rapports du LIST.

Khalfan, Ashfaq, Astrid Nilsson Lewis, Carlos Aguilar, Jacqueline Persson, Max Lawson, Nafkote Dabi, Safa Jayoussi, and Sunil Acharya. 2023. "Climate Equality: A planet for the 99%."

Kormos, Christine, and Robert Gifford. 2014. "The validity of self-report measures of proenvironmental behavior: A meta-analytic review." Journal of Environmental Psychology, 40: 359–371.

Kormos, Christine, Robert Gifford, and Erinn Brown. 2015. "The influence of descriptive social norm information on sustainable transportation behavior: A field experiment." Environment and Behavior, 47(5): 479–501.

Krupka, Erin L, and Roberto A Weber. 2013. "Identifying social norms using coordination games: Why does dictator game sharing vary?" Journal of the European Economic Association, 11(3): 495–524.

Leviston, Zoe, Iain Walker, and S Morwinski. 2013. "Your opinion on climate change might not be as common as you think." Nature Climate Change, 3(4): 334–337.

Lindström, Björn, Simon Jangard, Ida Selbing, and Andreas Olsson. 2018. "The role of a "common is moral" heuristic in the stability and change of moral norms." Journal of Experimental Psychology: General, 147(2): 228.

MIMCO Capital. 2023. "Luxembourg ranks 2nd highest average annual salary in the world with USD 65.449 in 2018." MIMCO Capital News. https://www.mimcocapital.com/en/2023/09/22/luxembourg-ranks-2nd-highest-average-annual-salary-in-the-world-with-usd-65-449-in-2018/.

Mullen, Brian, Jennifer L Atkins, Debbie S Champion, Cecelia Edwards, Dana Hardy, John E Story, and Mary Vanderklok. 1985. "The false consensus effect: A meta-analysis of 115 hypothesis tests." Journal of Experimental Social Psychology, 21(3): 262–283.

Nolan, Jessica M, P Wesley Schultz, Robert B Cialdini, Noah J Goldstein, and Vladas Griskevicius. 2008. "Normative social influence is underdetected." Personality and social psychology bulletin, 34(7): 913–923.

Prentice, Deborah A, and Dale T Miller. 1996. "Pluralistic ignorance and the perpetuation of social norms by unwitting actors." In Advances in experimental social psychology. Vol. 28, 161–209. Elsevier.

Reese, Gerhard, Kristina Loew, and Georges Steffgen. 2014. "A towel less: Social norms enhance pro-environmental behavior in hotels." The Journal of Social Psychology, 154(2): 97–100.

Reisinger, James. 2022. "Subjective well-being and social desirability." Journal of public economics, 214: 104745.

Ross, Lee, David Greene, and Pamela House. 1977. "The "false consensus effect": An egocentric bias in social perception and attribution processes." Journal of experimental social psychology, 13(3): 279–301.

Saracevic, Selma, and Bodo B Schlegelmilch. 2021. "The impact of social norms on pro-environmental behavior: a systematic literature review of the role of culture and self-construal." Sustainability, 13(9): 5156.

Schultz, P Wesley. 1999. "Changing behavior with normative feedback interventions: A field experiment on curbside recycling." Basic and applied social psychology, 21(1): 25–36.

Schultz, Wesley P, Azar M Khazian, and Adam C Zaleski. 2008. "Using normative social influence to promote conservation among hotel guests." Social influence, 3(1): 4–23.

Shamir, Jacob, and Michal Shamir. 1997. "Pluralistic ignorance across issues and over time: Information cues and biases." Public Opinion Quarterly, 227–260.

Sokoloski, Rebecca, Ezra M Markowitz, and David Bidwell. 2018. "Public estimates of support for offshore wind energy: False consensus, pluralistic ignorance, and partisan effects." Energy Policy, 112: 45–55.

Sparkman, Gregg, Nathan Geiger, and Elke U Weber. 2022. "Americans experience a false social reality by underestimating popular climate policy support by nearly half." Nature communications, 13(1): 4779.

Swim, Janet K, and Nathaniel Geiger. 2021. "Policy attributes, perceived impacts, and climate change policy preferences." Journal of Environmental Psychology, 77: 101673.

Verheyden, Bertrand, Michel Tenikue, Philippe Van Kerm, Ángela Jiang-Wang, Francesco Fallucchi, and David Cristelo. 2024. "Driving Behavioral Change for an Economic and Social Transition towards more Resilience and Sustainability in Luxembourg: SOC2050." Rapports du LISER.

Wynes, Seth, Steven J Davis, Mitchell Dickau, Susan Ly, Edward Maibach, Joeri Rogelj, Kirsten Zickfeld, and H Damon Matthews. 2024. "Perceptions of carbon dioxide emission reductions and future warming among climate experts." Communications Earth & Environment, 5(1): 498.

Appendix

A2.1 Pre-registered analyses

Our pre-registered analyses are:

- Result 1a. Assessing the underestimation of behaviors and personal norms (comparing empirical expectations with mean behaviors and normative expectations with mean personal norms).
- Result 1b. Assessing the underestimation of policy support (comparing expected policy support with actual policy support).
- Result 3. Assessing the extent personal norms and social expectations predict behavior and policy support.
- Result 4. Assessing the extent social expectations predict personal norms.

Since in the pre-registration we did not indicate that we would use weights in our analyses, we replicate all our analyses with unweighted data in Appendices A2.2 and A2.3.

A2.2 Robustness check (I): Wilcoxon signed-rank tests and effect sizes with unweighted data

Table A2.1: Underestimation of sustainable behaviors and norms (unweighted)

	Own's answer (mean)	Expectation others (mean)	p-value	Cohen's d
Behaviors				
Vegetarian meals	14.38 meals	12.81 meals	0.000	0.42
Lower home temperature	9.92°	9.11°	0.000	0.56
Public mobility time	44.82 min	42.32 min	0.001	0.15
Donation of lottery gains	30.28%	28.50%	0.1	0.06
Personal Norms				
Vegetarian meals	16.39 meals	15.74 meals	0.000	0.23
Lower home temperature	9.71°	9.39°	0.000	0.21
Public mobility time	56.76 min	54.53 min	0.000	0.14

Note: p-values from Wilcoxon signed-rank tests.

Table A2.2: Underestimation of support for restrictive and taxing policies (unweighted)

	Supporters (%)	Expected supporters (%)	p-value	Cohen's d
Regulation on red meat	61.61	35.26	0.000	0.72
VAT on meat	37.31	24.23	0.000	0.36
Fossil fuel rationing	51.16	35.87	0.000	0.40
Rental tax on poor insulation	64.40	44.83	0.000	0.52
Ban on cars in city center	50.31	33.80	0.000	0.43
Toll on highways	21.67	19.23	0.000	0.08

Note: p-values from Wilcoxon signed-rank tests.

A2.3 Robustness check (II): pre-registered OLS regressions without controls and with unweighted data

Table A2.3: Do social expectations predict behaviors?

	Vegetarian meals		Lower home temperature		Public mobility time		Donation of lottery gains
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Weighted data without controls							
Empirical expectation	0.425***		0.409***		0.495***		0.938***
Normative expectation		0.268***		0.353***		0.237***	
Panel B: Unweighted data with controls							
Empirical expectation	0.451***		0.331***		0.485***		0.878***
Normative expectation		0.296***		0.322***		0.196***	
Panel C: Unweighted data without controls							
Empirical expectation	0.446***		0.333***		0.486***		0.889***
Normative expectation		0.298***		0.312***		0.197***	
Observations	1292	1292	1292	1292	1292	1292	1292

* p<0.05, ** p<0.01, *** p<0.001.

Results from OLS regressions. Coefficients in Columns 1-6 are expressed as SD. Coefficient in Column 7 is expressed as percentage points.

Table A2.4: Do personal norms predict behaviors?

	Vegetarian meals	Lower home temperature	Public mobility time
	(1)	(2)	(3)
Panel A: Weighted data without controls			
Personal norms	0.591***	0.563***	0.292***
Panel B: Unweighted data with controls			
Personal norms	0.561***	0.552***	0.257***
Panel C: Unweighted data without controls			
Personal norms	0.570***	0.547***	0.258***
Observations	1292	1292	1292

* p<0.05, ** p<0.01, *** p<0.001

Results from OLS regressions. Coefficients are expressed as SD.

Table A2.5: Does expected policy support predict one's own support for policies?

	Regulation on red meat (1)	VAT on meat (2)	Fossil fuel rationing (3)	Rental tax on poor insulation (4)	Ban on cars in city center (5)	Toll on highways (6)
Panel A: Weighted data without controls						
Expected support	0.721***	0.806***	1.026***	0.744***	0.941***	0.888***
Panel B: Unweighted data with controls						
Expected support	0.701***	0.808***	1.014***	0.753***	0.946***	0.838***
Panel C: Unweighted data without controls						
Expected support	0.715***	0.841***	0.995***	0.744***	0.936***	0.856***
Observations	1292	1292	1292	1292	1292	1292

* p<0.05, ** p<0.01, *** p<0.001.

Results from OLS regressions. Coefficients expressed as percentage points.

Table A2.6: Do social expectations predict personal norms?

	Vegetarian meals		Lower home temperature		Public mobility time	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Weighted data without controls						
Empirical expectation	0.363***		0.403***		0.427***	
Normative expectation		0.559***		0.699***		0.798***
Panel B: Unweighted data with controls						
Empirical expectation	0.378***		0.344***		0.340***	
Normative expectation		0.605***		0.667***		0.774***
Panel C: Unweighted data without controls						
Empirical expectation	0.372***		0.354***		0.345***	
Normative expectation		0.604***		0.671***		0.778***
Observations	1292	1292	1292	1292	1292	1292

* p<0.05, ** p<0.01, *** p<0.001.

Results from OLS regressions. Coefficients are expressed as SD.

A2.4 Exploratory OLS regressions

Table A2.7: Underestimation of behaviors (heterogeneity by one's behavior)

	Vegetarian meals		Lower home temperature		Public mobility time		Donation of lottery gains
	Empirical	Normative	Empirical	Normative	Empirical	Normative	Empirical
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
> mean	2.850***	1.318***	0.871***	0.792***	10.36***	6.013***	0.208***
Constant	-2.422***	-0.681*	-1.249***	-0.667***	-8.279***	-2.729	-0.0860***
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1292	1292	1292	1292	1292	1292	1292

* p<0.05, ** p<0.01, *** p<0.001

Results from OLS regressions with weighted data. > mean = 1 if one's behavior is above average, and 0 if one's behavior is below average. Dependent variable is underestimation of behaviors and norms (expectation - mean response).

Table A2.8: Underestimation of policy support (heterogeneity by one's support)

	Regulation on red meat	VAT on meat	Fossil fuel rationing	Rental tax on poor insulation	Ban on cars in city center	Toll on highways
	(1)	(2)	(3)	(4)	(5)	(6)
Support	0.105***	0.101***	0.180***	0.188***	0.186***	0.166***
Constant	-0.316***	-0.169***	-0.227***	-0.280***	-0.247***	-0.0428**
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1292	1292	1292	1292	1292	1292

* p<0.05, ** p<0.01, *** p<0.001

Results from OLS regressions with weighted data. Coefficients expressed as percentage points. Support = 1 if a policy is supported, and 0 otherwise. Dependent variable is underestimation of policy support (expected supporters - actual supporters).

A2.5 Questionnaire instructions

This appendix presents the instructions of our main outcomes of interest, control variables and incentivization tasks.

A2.5.1 Demographics

What is your age?

- 18-25
- 26-35
- 36-50
- 51-65
- 66+

What gender do you identify with?
[Male / Female / Other (write) / Prefer not to say]

Which is the highest level of education that you have completed?

- Elementary school (primary school)
- Secondary (first or second cycle of the secondary education)
- Post-secondary but non-tertiary education (professional school or preparatory classes to tertiary education)
- Tertiary (university or other higher education degree).

What was the range of your **monthly net household's income** in September 2022?
With "monthly net household's income" we mean the total of net income of all persons in your household, but also other monthly revenues such as family allowances, unemployment benefits, pensions, and any other sources.

- 0 - 1,250 euros
- 1,250 - 2,000 euros
- 2,000 - 4,000 euros
- 4,000 - 6,000 euros
- 6,000 - 8,000 euros
- 8,000 - 12,500 euros
- Greater than 12,500 euros
- I prefer not to say

A2.5.2 Attention check question

This is a question for us to check that you are paying attention to the survey. Please let us know that you are reading all our questions carefully by selecting "1" on the response scale.

[Slider ranging from 1 to 10]

Only participants who selected "1" on both attention check questions were allowed to continue the survey.

A2.5.3 Bonus payment instructions

When we asked you whether you frequently pay attention to a list of specific actions (such as to avoid taking the plane), we asked you right after to guess how other respondents behave.
We will ask you **10 guessing questions** similar to this one, but from now on you can earn a **bonus depending on the quality of your guesses**. Hence to succeed in these guessing games, you need to focus on what you think are the **other** respondents' actions and opinions.

Would you like more details on the bonus scoring? (click)

For these guessing games, the closer your guess will be to the most frequent answer of the other respondents, the more points you will earn. For each of these 10 guessing games, you can earn up to 10 points. Hence the total survey score is maximum 100. Your bonus will depend on how your total score is ranked relative to all the other participants. More specifically, your bonus will be:

- **an additional 30€** if your score is among the **50 best scores**,
- **an additional 20€** if your score is ranked between the **51st and 100th** position,
- **an additional 10€** if your score is ranked between the **101st and 400th** position,

Note that, as we aim at getting 1500 participants in this survey, you have more than 25% chance of receiving an extra bonus.

A2.5.4 Animal proteins

In the following set of questions, we will talk about consumption of **meat, fish or seafood**. By "one meal containing meat, fish or seafood", we mean a meal containing at least 50 grams of it.

Self-reported behavior

Over the last 7 days, **how many meals** containing meat, fish or seafood did you eat?

[Slider ranging from 0 to 21]

Empirical expectation

Guessing game
Please guess what will be the most frequent answer in the survey to the question "Over the last 7 days, **how many meals** containing meat, fish or seafood did you eat?"

[Slider ranging from 0 to 21]

Interested to know how your score is calculated? (click)

You will obtain a score of 10 if you find the correct answer, and will lose one point for each unit that separates your answer from the correct answer. Your score cannot go below 0.

- Example 1: if your guess is 10 meals but the most frequent answer is 15 meals, the gap is 5 units, so you will earn 10-5=5 points.
- Example 2: if your guess is 5 meals but the actual most frequent answer is 20 meals, the gap is more than 10 units, so you will earn 0 points.

Personal norm

Due to the amount of energy needed to produce them, **meat, fish and seafood are not a sustainable source of proteins**. With these sustainability concerns in mind, how many meals per week containing meat, fish or seafood do you think it would be **ethically appropriate** to eat?

[Slider ranging from 0 to 21]

Normative expectation

Guessing game
Please guess what will be the most frequent answer in the survey to the question "With these sustainability concerns in mind, how many meals per week containing meat, fish or seafood do you think it would be **ethically appropriate** to eat?".

[Slider ranging from 0 to 21]

Interested to know how your score is calculated? (click)

You will obtain a score of 10 if you find the correct answer, and will lose one point for each unit that separates your answer from the correct answer. Your score cannot go below 0.

- Example 1: if your guess is 10 meals but the most frequent answer is 15 meals, the gap is 5 units, so you will earn 10-5=5 points.
- Example 2: if your guess is 5 meals but the actual most frequent answer is 20 meals, the gap is more than 10 units, so you will earn 0 points.

Policy support

With these sustainability concerns in mind, let us **assume (hypothetically)** that by 2025, two new important measures are considered:

- reducing the **available quantity of red meat** by imposing strict regulations (e.g. only allowing the sale of meat from cull cows that have reached the end of their milk production or breeding function, or from farms that adopt strict environmental protocols)
- increase the **Value Added Tax rate (VAT)** on meat, fish or seafood to **17%** (current rate is 3%)

With the money obtained from this tax on meat, and to compensate the reduced availability of red meat, **a massive plan to promote vegetarian alternatives** would be put in place.

Here are some examples of measures that can be implemented under this plan (click)

- Vegetarian alternatives are made more visible than meat in supermarkets stall
- Vegetarian alternatives are subsidized (price reduced by half)
- A label comparing the environmental impact per protein intake of meat and of vegetarian alternatives
- A label comparing the price per protein intake of meat and of vegetarian alternatives
- Free consultations with nutritionists to ensure that vegetarian meals contain all necessary nutrients

Under this massive meat substitution plan...
Would you support the **reduction** of the available quantity of **red meat** in supermarket stalls, butcheries and restaurants?
[Yes/No]
Would you support an increase of the **Value Added Tax rate (VAT)** on meat, fish or seafood to **17%**?
[Yes/No]

Policy support (expectation)

Guessing game
If a massive meat substitution plan is implemented, out of 100 participants, how many do you think would...

- support regulations **reducing** the availability of **red meat**?
- support the **VAT increase on meat, fish or seafood**?

Interested to know how your score is calculated? (click)
For each of the two questions, you will obtain a score of 5 if you find the correct answer, and will lose one point for each unit that separates your answer from the correct answer. Your score cannot go below 0.

A2.5.5 Home temperature

In the following set of questions, we will ask about your housing.

Self-reported behavior

What is the **usual temperature** of your dwelling when you are at home when it is less than 10 degrees outside? (in degree Celsius)

[Slider ranging from 15 to 30]

Empirical expectation

Guessing game

What do you think is the most frequent answer to the question: "What is the **usual temperature** of your dwelling when you are at home when it is less than 10 degrees outside?" ?

[Slider ranging from 15 to 30]

Interested to know how your score is calculated? (click)

You will obtain a score of 10 if you find the correct answer, and will lose two points for each °C that separates your answer from the correct answer. Your score cannot go below 0.

Personal norm

Heating one's home with **fossil energy sources (e.g., oil, coal and natural gas) is not sustainable** in the long term.

From a sustainability viewpoint, what do you think is the maximal temperature that it is **ethically appropriate** to set in...

- a well-insulated dwelling when it is less than 10 degrees outside?
- a badly-insulated dwelling when it is less than 10 degrees outside?

[Slider ranging from 15 to 30]

Normative expectation

Guessing game

What do you think are the most frequent answers to the questions

"From a sustainability viewpoint, what do you think is the maximal temperature that it is **ethically appropriate** to set in..."

- a well-insulated dwelling when it is less than 10 degrees outside?
- a badly-insulated dwelling when it is less than 10 degrees outside?

[Slider ranging from 15 to 30]

Interested to know how your score is calculated? (click)

For each of the two questions, you will obtain a score of 5 if you find the correct answer, and will lose one point for each °C that separates your answer from the correct answer. Your score cannot go below 0.

Policy support

With these sustainability concerns in mind, let us **assume (hypothetically)** that by 2025, two new important measures are considered:

- A **rationing on fossil energy** sources, based on household size and needs (for instance, a quota on heating oil or gas, depending on the size of the household)
- A **tax** of 10% on the rental income perceived by **landlords** who rent out **accommodations with insufficient energy efficiency** (EPC rating of D or worse).

These two policies would facilitate the implementation of a **massive housing plan to improve energy efficiency**.

Here are some examples of measures that can be implemented under this plan (click)

- short-term support to reduce energy consumption
- free heating system servicing (tune-up and maintenance inspection)
- personalized support (at home or online) to change habits and reduce consumption
- support to structural renovation and equipment investments
- free energy audits to determine ECT rating and identify necessary investments
- need-based subsidies on insulation investments and renewable energy equipment
- zero interest rates on renovation and energy-efficient equipment investments

Under this massive housing plan...
Would you support the **rationing on fossil energy** sources?
[Yes/No]
Would you support a **tax** of 10% on the rental income perceived by **landlords** who rent out **accommodations with insufficient energy efficiency**?
[Yes/No]

Policy support (expectation)

Guessing game
If a massive housing plan is implemented, out of 100 participants, how many do you think would...

- Support the **rationing on fossil energy** sources?
- Support the **tax** paid by **landlords** who rent out **accommodations with insufficient energy efficiency**?

Interested to know how your score is calculated? (click)
For each of the two questions, you will obtain a score of 5 if you find the correct answer, and will lose one point for each unit that separates your answer from the correct answer. Your score cannot go below 0.

A2.5.6 Mobility

In the following set of questions, we will ask about your mobility.

Self-reported behavior

Let us consider a hypothetical trip that you would regularly take and which, if done by car, would take about 30 minutes (for example going to work).
Now consider that there exists for this trip an **alternative mobility solution**: public transportation (tram, train, bus), possibly in combination with an alternative transport mode (scooter, bike, walking). This alternative solution may of course take a different amount of time.
Compared to the 30-minute car trip option, what would be the maximal trip duration that you would accept if you were to systematically use public transport instead?
I would systematically use public transport if it took (in minutes) at most...

[text box to write]

How frequently do you use public transport (bus, train, tram)?

- Daily
- Several times a week
- Several times a month
- Several times a year
- Never

How frequently do you **walk, cycle** or use a (electric) **scooter** for a trip of 10 minutes or more?

- Daily
- Several times a week
- Several times a month
- Several times a year
- Never

Empirical expectation

Guessing game

What do you think is the most frequent answer given among all the other participants to the question: "Compared to the 30-minute car trip option, what would be the maximal trip duration that you would accept if you were to systematically use public transport instead?"

I think that other respondents most frequently state that they would systematically use public transport if it took (in minutes) at most...

[text box to write]

Interested to know how your score is calculated? (click)

You will obtain a score of 10 if you guess correctly the most frequent answer, and will lose one point for each minute that separates your guess from the most frequent answer. Your score cannot go below 0.

- Example 1: if your guess is 30 minutes but the most frequent answer is 35 minutes, the gap is 5 minutes, so you will earn 10-5=5 points.
- Example 2: if your guess is 30 minutes but the actual most frequent answer is 15 minutes, the gap is more than 10 minutes, so you will earn 0 points.

Personal norm

Let us still consider this regular 30-minute trip by car. It is well established that **public transport and soft mobility** have a significantly **smaller impact on the environment** than the car. From this perspective, it might be justified to use public transport despite spending more time.

In your opinion, what is the maximal **additional time** (on top of the 30 minutes) that would be **ethically appropriate** to concede when switching to public transport?

"In view of adopting a more sustainable lifestyle, I think that it would be **ethically appropriate** to systematically switch to public transport as long as the **additional time** (in minutes, on top of the initial 30 minutes) does not exceed ... "

[text box to write]

Normative expectation

Guessing game

What do you think is the most frequent answer given by the other participants when they are asked: "What is the maximal additional time (on top of the 30 minutes) that it would be **ethically appropriate** to concede when switching to public transport?"

I think that the most frequent answer (in minutes, on top of the initial 30 minutes) is...

[text box to write]

Interested to know how your score is calculated? (click)

You will obtain a score of 10 if you guess correctly the most frequent answer, and will lose one point for each minute that separates your guess from the most frequent answer. Your score cannot go below 0.

- Example 1: if your guess is 30 minutes but the most frequent answer is 35 minutes, the gap is 5 minutes, so you will earn 10-5=5 points.
- Example 2: if your guess is 30 minutes but the actual most frequent answer is 15 minutes, the gap is more than 10 minutes, so you will earn 0 points.

Policy support

With these sustainability concerns in mind, let us **assume (hypothetically)** that by 2025, two new important measures are considered:

- to **ban all cars** in densely populated areas
- a **5-euro toll** every time one uses Luxembourg's motorways (i.e. 10 euros per day on a two-way commute for example)

These two policies would facilitate the implementation of a **massive mobility plan based on public transport and soft mobility**.

Here are some examples of measures that can be implemented under this plan (click)

- Giving higher priority to buses in bottlenecks
- Developing the soft mobility infrastructure (e.g., more bike lanes, more rental bikes,...)
- Reduce delays / uncertainty about trip duration
- Reduce the duration of trips (fewer connections, provide more public transport hubs)
- Make hubs and public transport safer
- Increase provision in order to prevent overcrowding (increase frequency of services, expand hours of services early morning and late evening)

Under this massive mobility plan...
Would you support the **car ban** in densely populated areas?
[Yes/No]
Would you support the **5-euro toll** on Luxembourg's motorways?
[Yes/No]

Policy support (expectation)

Guessing game

If a massive public mobility plan is implemented, out of 100 participants, how many do you think would...

- Support the **car ban** in densely populated areas?
- Support the **5-euro toll** on Luxembourg's motorways?

Interested to know how your score is calculated? (click)

For each of the two questions, you will obtain a score of 5 if you find the correct answer, and will lose one point for each unit that separates your answer from the correct answer. Your score cannot go below 0.

A2.5.7 Monetary donations

Donation of bonus earnings

Thank you for having answered this survey. Now let us talk about your reward.

Your reward (fixed participation fee plus the bonuses from guessing games) will be sent to you by Amazon vouchers. You have, however, the option to make a different use of this reward.

To allow individuals and companies to compensate their CO2 emissions, there is a framework for "carbon offsetting" via a certified market for voluntary "**carbon credits**". Want to know how it works? (click)

Individuals and companies can offset their CO2 emissions by funding certified climate action projects (reforestation, public investments in renewable energy sources,...). To achieve this, individuals can purchase "carbon credits" which are issued on the Voluntary Carbon Market" (VCM) by government bodies and NGO's. These VCM projects need to obtain certification that they indeed reduce greenhouse gas emissions. This certification is ensured by independent standards organizations, such as the world leader VCS (Verified Carbon Standard).

Would you agree to renounce your earnings and have them allocated to the **purchase of carbon offsets**? LISER will collaborate with a Luxembourg-based organization which exclusively proposes green projects that were audited by third parties and that received the highest international carbon-offsetting standards.

Your personal decision will not be communicated in any way. Only the total amount of the gains conceded by all participants will be disclosed.

[Yes, I renounce my earnings / No, I want to keep my earnings]

Donation of lottery gains

In addition to the guaranteed payments, **10 participants** will be randomly selected and each will earn an extra **250 euros** (lottery rules available on LISER's webpage). However, we offer the possibility of making an alternative use of the lottery gains.

With 250€ you might offset 4 months of carbon emissions, and if you give up the 250€ in each wave of the survey, you might offset a whole year of carbon emissions. Want to know how this is calculated? (click)

Offsetting the annual CO2 emissions of an average Luxembourg citizen costs about 750€. This figure is calculated as follows: In October 2022, offsetting 1 ton of CO2 emissions in the European Union costs about €60 (source: Ember Climate). The Ministry of Environment of Luxembourg recently declared that 8.08 million tons of CO2-equivalent were emitted by Luxembourg in 2021. This makes an emission of 12.6 tons per inhabitant.

In this wave, **how much of the 250€ would you like to be allocated to the purchase carbon credits**? If you win the lottery, we will allocate the money according to your decision and provide you the remaining part in the form of an Amazon voucher.

Note: your chances of winning the lottery are not affected in any way by your responses to the survey.

[Slider thanging from 0 - All to myself to 250 - All to carbon credits]

Out of the 250€, how much do you think the **other participants** would donate for carbon offsetting on average?

[Slider thanging from 0 - All to myself to 250 - All to carbon credits]

Chapter 3

Effectiveness of correcting misperceived social norms: A longitudinal experimental approach with sustainable behaviors and policy support

Effectiveness of correcting misperceived social norms: A longitudinal experimental approach with sustainable behaviors and policy support

with Francesco Fallucchi (University of Bergamo), Philippe Van Kerm (University of Luxembourg), and Bertrand Verheyden (Luxembourg Institute of Socio-Economic Research)

3.1 Introduction

Household consumption, particularly in the areas of food, housing, and transportation, accounts for nearly three-quarters of global carbon emissions. There is broad consensus that systematic and sustained behavioral change in these domains, along with policies targeting household consumption, is essential for transitioning to a low-carbon society (Druckman and Jackson 2016, Dubois *et al.* 2019, Hertwich and Peters 2009). This is especially true in high-income countries, which are not only the most polluting (Khalfan *et al.* 2023), but also where individual willingness to act on climate change is lowest (Andre *et al.* 2024a).

Misperceived social norms have been identified as a common barrier to climate action (Andre *et al.* 2024a,b, Sparkman, Geiger and Weber 2022, Geiger and Swim 2016, Sokoloski, Markowitz and Bidwell 2018). As shown in past research and discussed in Chapter 2, these misperceptions are often driven by cognitive biases such as pluralistic ignorance, which occurs when individuals misperceive the prevalence of an opinion or behavior (Prentice and Miller 1996), and the false consensus effect, which occurs when individuals overestimate the prevalence of their own opinion or behavior (Ross, Greene and House 1977). These biases can have significant consequences. A large body of research shows that our preferences are often conditional on our social expectations (Bicchieri 2005, 2016), and this applies to sustainable behaviors as well (Cialdini and Jacobson 2021, Saracevic and Schlegelmilch 2021). Misperceived social norms can therefore be harmful because they lead individuals to conform to undesirable norms. However, research has also shown that correcting these misperceptions can have positive effects. For instance, revealing the true extent to which others are concerned about climate change and support efforts to address it has been shown to increase individuals' willingness to discuss climate change with others, donate money, and support climate policies (Geiger and Swim 2016, Andre *et al.* 2024b).

In this paper, we examine whether correcting misperceptions about the extent to which others engage in everyday sustainable consumption, the level of engagement they consider appropriate, and their support for policies that restrict unsustainable behaviors can promote participants' own

sustainable behaviors, personal norms, and policy support.[1] As in Chapter 2, we focus on the three behavioral domains with the highest potential for carbon savings: increasing vegetarian consumption, reducing home heating, and using public or soft mobility options (Hitaj, Igos and Gibon 2022). We also investigate whether the effects of these interventions persist over the longer term. In Luxembourg, adopting a vegetarian diet, reducing home heating, and commuting by public transport instead of by car can save up to 4.8 tonnes of CO_2eq per year, approximately 37% of the 13 tonnes emitted annually by the average Luxembourger (Hitaj, Igos and Gibon 2022). Thus, sustained behavioral change in these domains, supported by both individual habit formation and structural policies targeting personal behavior, can yield significant environmental benefits.

We conducted a three-wave longitudinal survey between November 2022 and July 2023, with each wave spaced three months apart, involving 912 adult participants from Luxembourg. In Wave 1, we measured participants' baseline behaviors and empirical expectations by asking about their weekly frequency of vegetarian consumption, home temperature when it is cold outside, and use of public or soft transport. We then asked them to estimate the same behaviors for other participants. To measure baseline personal norms and elicit normative expectations, we asked participants which levels of these behaviors they considered ethically appropriate from a sustainability perspective, and then asked them to estimate the most common response given by other participants to that same question. Participants were also asked whether they would support six hypothetical policies aimed at restricting or taxing unsustainable behaviors across the three behavioral domains. For each domain, we included one tax and one regulatory restriction. Finally, participants were asked to guess the proportion of others in the sample who supported each policy. As described in Chapter 2, all second-order beliefs were elicited in an incentive-compatible manner. Our findings revealed that participants systematically and substantially underestimated others' sustainable behaviors, norms, and policy support across all domains. Consistent with previous literature, these results reflected both pluralistic ignorance and the false consensus effect. We also found that social expectations were strong predictors of individual behaviors, personal norms, and policy support, underscoring the potential of correcting misperceptions to promote behavioral change. These results are reported in detail in Chapter 2 of this dissertation.

Following these findings, we introduced an experiment in Wave 2. Participants were randomly assigned to a Control group (n = 305), a Norms treatment group (n = 304), or a Policy treatment

[1]Although correcting misperceptions about concern for climate change has proven effective in increasing support for climate policies targeting businesses and institutions, policies that target individual behavior are generally less popular and more costly to support (Swim and Geiger 2021). Therefore, it remains an important question whether the positive effects of correcting misperceptions extend to such outcomes.

group (n = 303). In the Norms treatment, we revealed to participants the actual levels of behavior and personal norms reported by the majority in Wave 1. In the Policy treatment, we revealed to them the actual share of participants who supported each of the six policies in Wave 1. Participants in the Control group received no information about Wave 1 responses. We chose to present both descriptive norms (what others do) and injunctive norms (what others find appropriate) in the Norms treatment because previous research shows that combining these messages is more effective than presenting either alone (Hallsworth *et al.* 2017, Schultz, Khazian and Zaleski 2008). Moreover, experts recommend changing both empirical and normative expectations to achieve successful behavioral change (Bicchieri 2016).

After exposure to the treatments (or lack thereof, in the case of the Control group), we asked participants to report again their personal norms, their intended behavior levels for the following three months, and their support for the six policies, using the same measures as in Wave 1. Three months later, in Wave 3, participants reported their actual behavior levels, allowing us to assess whether they had followed through on their intentions. Home heating was excluded from this assessment, as Wave 3 took place in early summer. We also asked participants again about their policy support. This design enabled us to examine whether the treatments had lasting effects or only a short-term impact.

We collected the same post-treatment outcomes across all groups, regardless of treatment assignment. This allowed us to estimate the effects of both targeted and non-targeted treatments. We define targeted treatments as those that correct misperceptions directly related to the outcome (for example, the Policy treatment is targeted at policy support), and non-targeted treatments as those that correct misperceptions about a different outcome (for example, the Norms treatment is non-targeted for policy support). A recurring question in the literature is whether sustainable behavior and support for sustainable policies act as complements or substitutes. Some studies have found that interventions promoting sustainable behaviors can crowd out support for climate policies, possibly by shifting attention or perceived responsibility from policy to personal action (Werfel 2017). Other studies, however, find no evidence of such negative spillover effects (Maki *et al.* 2019, Sparkman, Attari and Weber 2021). By examining the effects of both targeted and non-targeted treatments on our outcomes, we can assess whether our treatments could potentially crowd out the non-targeted outcomes.

Controlling for the baseline values of our post-treatment outcomes measured in Wave 1, we found that our targeted treatments immediately increased participants' sustainable personal norms by an average of 0.25 SD, intended levels of sustainable behavior by an average of 0.15 SD, and policy

support for regulatory policies (those revealed to be supported by a majority) by an average of 11 percentage points. However, this was true only for the behavioral domains of vegetarian consumption and home heating. Participants showed no response in behavior and policy support in the domain of public and soft mobility, and the treatments actually shifted personal norms in a less sustainable direction (-0.22 SD). Participants were also unresponsive to the treatments in their support for taxing policies. When asked three months later, the positive effects of the targeted treatments on actual behavior (0.15 SD) and policy support (10 percentage points) persisted for vegetarian consumption. We found no evidence of negative spillovers from the non-targeted treatments. In fact, we either observed no spillover effects or positive spillover effects, providing support for the idea that interventions to increase sustainable behaviors and interventions to increase policy support do not act as substitutes in promoting behavioral change.

We also examined whether our treatments had heterogeneous effects depending on participants' prior beliefs. Since we expect the treatments to operate through positive belief updating, they should be particularly effective for individuals who initially held pessimistic beliefs about others and subsequently revised those beliefs upward. For example, Andre *et al.* (2024b) found that correcting misperceptions about others' willingness to fight climate change increased monetary donations only among individuals with pessimistic prior beliefs. If people tend to conform to their perceived norms (Prentice and Miller 1993, Bicchieri 2005), then behavior change should be most likely among those who had underestimated a sustainable norm and needed to update their beliefs positively. To test this, we replicated our main analyses separately for participants who held pessimistic beliefs about a given personal norm, behavior, or policy support in Wave 1 and for those who did not.

Similarly, we aimed to examine treatment heterogeneity based on prior behavior. If individuals change their behavior to conform to the new norm, treatments should be particularly effective for those who behaved below the norm and need to make an upward behavior adjustment. Therefore, we replicated our analyses separately for participants who reported a sustainable behavior level above the average (whom we refer to as sustainable individuals) and those who reported a level below the average (whom we refer to as unsustainable individuals) in Wave 1. For policy support, we conducted the analyses separately for participants who supported a given policy and those who did not.

The main advantage of conducting our analyses separately based on prior beliefs and behaviors is that we can assess not only whether the treatments were equally effective across different subgroups but also whether they backfired for any of them. If individuals tend to conform to perceived norms, treatments may backfire among those who held optimistic beliefs (and thus revise their beliefs

negatively) or among those who behaved above the norm (and thus need to adjust their behavior downward to align with the norm). For example, in a study on household energy conservation, descriptive norm information prompted high-consuming households to reduce consumption, but also caused low-consuming households to increase theirs (Schultz, Khazian and Zaleski 2008). This kind of asymmetric effect based on priors may help explain why norm-based interventions do not always succeed and can sometimes backfire (e.g. Griesoph *et al.* 2021, Richter, Thøgersen and Klöckner 2018, Gravert and Collentine 2021). Our heterogeneity analyses, therefore, help us identify whether our treatments triggered such unintended effects.

Indeed, our heterogeneity analyses showed that the targeted treatments were more likely to increase personal norms, sustainable behaviors, and policy support among participants with pessimistic beliefs and those who behaved unsustainably. These analyses also helped clarify the negative effect of the Norms treatment on personal norms related to public and soft mobility. This adverse effect appeared only among participants who held non-pessimistic beliefs and those who behaved above average. Although the treatments produced some unintended effects on personal norms, we did not find any evidence that they led to behavioral or policy support backfiring. The inclusion of both descriptive and injunctive norms in our treatments may have played a role in preventing such outcomes. For example, Schultz, Khazian and Zaleski (2008) showed that adding injunctive information to a descriptive norm message eliminated the negative response previously observed among low-energy consumers. Our results also align with the findings of Andre *et al.* (2024b), who reported no behavioral backfiring in monetary donations among individuals with optimistic prior beliefs about others' willingness to fight climate change. At the same time, our findings urge caution. Although correcting misperceptions can be an effective tool for promoting sustainable behavior and policy support, it may still lead to unintended negative effects in other areas for certain subgroups, such as personal norms.

Additionally, we found that positive spillover effects from non-targeted treatments were not more likely to occur among participants with prior pessimistic beliefs or unsustainable behaviors. This supports our expectation that targeted treatments operate specifically through positive belief-updating and upward behavior adjustment to conform to the norm, rather than affecting pessimistic or unsustainable participants per se, since non-targeted treatments did not correct participants' beliefs or provide norm-related information for the target outcome.

Finally, we observed that the positive effect of the treatments was consistent across personal norms, behaviors, and policy support in the domains of vegetarian consumption and home heating, especially for the first domain, where the effects persisted in the longer term. However, participants

showed complete inelasticity in behavior and policy support in the public and soft mobility domain, and even negative elasticity in personal norms. We tested three possible explanations for this:

1. *Can the lack of positive treatment effects on public and soft mobility be explained by the fact that participants, on average, were less pessimistic in this domain?* As described in Chapter 2, average misperceptions across all measures were consistently smaller for public and soft mobility. However, our heterogeneity analyses based on prior beliefs revealed that the lack of effectiveness cannot be attributed to the prevalence of positive belief updating, as treatments were no more effective for participants who held prior pessimistic beliefs in this domain.

2. *Can the lack of positive treatment effects be explained by the fact that participants, on average, were more sustainable in this domain?* Once again, heterogeneity analyses revealed that the lack of effectiveness cannot be explained by having fewer participants behaving below the norm or not supporting policies, as treatments were not more effective for the most unsustainable participants in this domain.

3. *Can the lack of positive treatment effects be explained by the fact that participants were less likely to remember the treatment content in this domain?* In Wave 3, we introduced a memory task to evaluate how well our participants remembered the content of our treatments introduced in Wave 2. However, once again, our results reveal that this cannot explain the lack of effectiveness.

Having ruled out these potential explanations, we propose a remaining plausible interpretation: correcting misperceptions may be less effective in changing behaviors in domains that are costlier to adopt. Mobility patterns depend heavily on individual constraints, such as access to reliable alternatives to cars, place of residence, and commuting needs. This may make behavior and support for policies restricting car usage more inelastic to policy interventions that do not address these structural differences. Indeed, despite the Luxembourgish government's implementation of strong economic incentives to reduce car use, such as making all public transportation within the country completely free of charge since 2020 (Rose 2023), car usage in Luxembourg remains among the highest in the EU (Heindrichs 2024). Hence, behavioral change in this domain may require more substantial structural support from institutions.

Our study contributes broadly to the literature on norm-based treatment interventions aimed at promoting sustainable behavior (Schultz, Khazian and Zaleski 2008, Nolan *et al.* 2008, Schultz 1999, Ferraro, Miranda and Price 2011, De Groot, Abrahamse and Jones 2013, Goldstein, Griskevicius

and Cialdini 2007, Reese, Loew and Steffgen 2014, Kormos, Gifford and Brown 2015, Sparkman *et al.* 2021). As many studies have shown, while norm-based interventions can be a promising policy avenue, they do not always work and can even backfire (Griesoph *et al.* 2021, Richter, Thøgersen and Klöckner 2018, Gravert and Collentine 2021). As noted by Bicchieri (2016), a proper diagnosis of the causal influencers of behavior in a population is necessary before designing interventions to promote behavioral change. We add sophistication to traditional norm-based interventions by first diagnosing the causes of barriers to behavioral change (such as underestimations of sustainable norms) and then designing targeted and customized interventions to address these barriers. Our study also shows that positive belief updating and the potential for upward behavioral adjustment toward the norm may be necessary conditions for the effectiveness of norm-based treatments, conditions that are much more likely to arise in situations of pluralistic ignorance or false consensus effects. In contrast, norm-based interventions may be less effective in the absence of these cognitive biases.

Our study also contributes to the literature on correcting misperceptions as a way to promote behavior, an increasingly popular intervention used in other areas such as female labor participation or support for partisan violence (Bursztyn, González and Yanagizawa-Drott 2020, Mernyk *et al.* 2022). We provide evidence that correcting misperceptions about others' specific behaviors, personal norms, and policy support can change participants' everyday sustainable behaviors, personal norms, and policy support, and these effects can last in the longer term. However, we also show that this effect may be moderated by factors such as the potential and direction of belief updating, the availability of room for upward behavioral adjustment in the community, and other constraints that may impede behavior change.

We contribute to the literature investigating spillover effects between sustainable behaviors and support for sustainable policies (Maki *et al.* 2019, Sparkman, Attari and Weber 2021, Werfel 2017, Truelove *et al.* 2014) by providing additional evidence that interventions promoting behaviors and interventions promoting policy support do not necessarily crowd out each other.

Finally, in line with past research and with preliminary analyses in Chapter 2, we establish causal evidence that individuals have conditional preferences in their sustainable behaviors and policy support (Vesely and Klöckner 2018, Andre *et al.* 2024a). Furthermore, consistent with our findings in Chapter 2, we also provide causal evidence that participants have conditional preferences in their personal norms, contributing to the literature on how different types of norms can impact and shape each other (Deutchman *et al.* 2024, Lindström *et al.* 2018). This is important because personal norms are strong predictors of behavior (Bašić and Verrina 2024). If correcting misperceptions can shift personal norms in a positive direction, it may lead to behavioral change through the internalization

of these updated norms.

The remainder of the paper is organized as follows. Section 3.2 describes our study design, experimental setup, and collected measures, and provides a description of our sample. Section 3.3 presents our model and results. Finally, Section 3.4 concludes.

3.2 Study Design

As described in Chapter 2, this chapter primarily draws on Waves 2 and 3 of the longitudinal survey study we conducted as part of a larger project in collaboration with the Ministry of the Economy in Luxembourg (Verheyden *et al.* 2024). We contacted 3,700 volunteers who had previously participated in LISER surveys and had consented to be recontacted for future research. We invited them to participate in a three-wave survey, with approximately three months between each wave: Wave 1 was launched in November 2022 (the design and findings of which are presented in Chapter 2 of this dissertation), Wave 2 in April 2023, and Wave 3 in July 2023. The volunteers were adult Luxembourgish residents and cross-border workers (individuals residing in neighboring countries who work in Luxembourg). A total of 912 participants completed all three waves of the study.

We followed the same protocol across all three surveys: participants were informed that their participation was voluntary and that they could withdraw at any time. They were assured that their responses would remain anonymous. Additionally, they were informed that the surveys contained two attention-check questions and that failing both would result in termination of the survey. The surveys were translated into English, French, German, and Portuguese. Respondents received a fixed reward of €10 for participating in each survey. In each wave, participants also had the opportunity to earn a bonus of €10 to €30 by ranking among the top 30% of scorers on questions that required them to guess the responses of other survey participants (these results are not discussed in this chapter). Furthermore, in each survey wave, 10 participants were randomly selected through a lottery and awarded a €250 prize. Full compensation for participation in the three waves was sent at the end of the third wave. Therefore, participants were compensated only once for their cumulative payment, which ranged between €30 and €120 (€10 guaranteed per wave plus a maximum bonus of €30 per wave), along with the potential €250 prize for each of the three lotteries. Importantly, before receiving their final compensation, participants had no information about their bonus or lottery earnings in each wave. These studies received IRB approval from LISER. Detailed information on the experiment instructions and the post-treatment measures collected in Waves 2 and 3 can be found in Appendices A3.5 and A3.6. Detailed information on the pre-treatment measures collected

in Wave 1 is provided in Appendix A2.5 of Chapter 2.

3.2.1 Pre-treatment measures

In Wave 1, we collected a broad set of demographic data on participants, including gender, age, income, and education. We then gathered data on self-reported sustainable behaviors, personal norms and policy support across three behavioral domains: vegetarian consumption, home heating reduction, and the use of public and soft mobility. We used this information to design the information treatments for Wave 2 and to establish the baseline values for our post-treatment outcomes. We also elicited participants' social expectations to measure their underestimations of others' sustainability. In other words, participants had to guess the most common responses given by all participants for each of the aforementioned measures, and they could receive an additional bonus of up to €30 if their guesses were accurate. Finally, we created two additional prior variables based on the data collected in Wave 1:

Prior behavior. We created a dummy variable that takes the value 1 if a participant reported a sustainable behavior level above the average in Wave 1 and 0 if they reported a level below the average. We use the terms "sustainable individual" and "unsustainable individual" throughout the text to refer to those individuals with above-average and below-average sustainable behavior in Wave 1. Note that this classification applies only to the specific behavior in question. For example, the same participant could be classified as sustainable in vegetarian consumption but unsustainable in home heating.

Prior beliefs. We created a dummy variable equal to 1 if a participant underestimated a given behavior, personal norm, or the share of policy supporters in Wave 1, and 0 otherwise. Specifically, we consider a participant to have underestimated a measure if they reported a social expectation lower than the mean response across all participants for a given behavior, personal norm, or share of supporters. We use the terms "pessimistic individual" and "non-pessimistic individual" throughout the text to refer to those who underestimated or did not underestimate a sustainable measure. As before, this classification is tied to the specific behavior, personal norm, or policy in question. For instance, the same participant could be classified as pessimistic about vegetarian consumption but not about home heating reduction.

Additional details of our pre-treatment measures are provided in Chapter 2.

3.2.2 Experimental treatments

In the second survey wave, we introduced an experiment featuring two information treatments. Participants were randomly assigned to either the Norms treatment (n = 304), the Policy treatment (n = 303), or the Control group (n = 305). The information treatments were presented before any post-treatment outcomes were measured. The remainder of the study was then identical for all participants.

Norms treatment. In the Norms treatment, we informed participants about both the actual and the appropriate behavior levels reported by the majority in the first wave, regarding vegetarian consumption, home heating reduction, and the use of public and soft transport. We began by reminding them that, in Wave 1, we had asked about their typical levels of consumption and the levels they considered "ethically appropriate," given the environmental impact, across three domains of everyday behavior: meat consumption, home heating, and car use.

We then explained that we had analyzed the data and would now inform them about the behaviors and attitudes of the majority. We clarified that “behaviors and attitudes of the majority” referred to those adopted by more than 50% of the population. Participants were also informed that the statistics were based on data from Wave 1 and had been adjusted to be representative of the Luxembourgish population in terms of age and gender.

Specifically, they learned that:

- The majority ate six or fewer meals containing meat, fish, or seafood per week and believed it would be ethically appropriate to eat four or fewer such meals per week.
- When the temperature outside was below 10°C, the majority kept their home temperature at 20°C or lower, and considered this level ethically appropriate. They were also informed that lowering home temperature by 1°C could reduce energy consumption by 3% to 10%.
- When faced with a 30-minute car trip, most people were willing to increase their travel time by 15 minutes or more to switch from using a car to public transport, and considered it ethically appropriate to increase travel time by 20 minutes or more. Additionally, participants were told that 69% of the sample used either public transport (bus, train, or tram) or engaged in soft mobility (e.g., walking or cycling for at least 10 minutes) several times a week or daily.

Policy treatment. In the Policy treatment, we informed participants about the actual share of supporters for each of the six policies presented in the first wave. We began by reminding them that, in Wave 1, we had asked questions about three areas of everyday life, meat consumption, housing,

and mobility. For each of these areas, we presented two hypothetical policies (a tax and a regulatory restriction) and asked whether they would support their introduction by 2025. We also reminded participants that the tax revenues from these policies would be used by the government to develop action plans offering alternative solutions to the population. We then informed them that, although respondents believed that none of the six policies would be supported by the majority, four out of the six were in fact supported by a majority. As in the Norms treatment, we explained that the statistics were based on data from Wave 1 and adjusted to be representative of the Luxembourgish population in terms of age and gender. Subsequently, participants learned that:

- 64% supported a tax of 10% on the rental income of landlords who rent out accommodations with poor energy efficiency
- 63% supported a strict regulation on red meat production
- 53% supported a rationing on fossil energy sources
- 50% supported a car ban in densely populated areas
- 37% supported a 17% VAT on meat
- 21% supported a 5-euro toll on Luxembourg's motorways

Control. The Control group did not receive any information about the responses from the first wave. They were given the same reminders as the Norms and Policy treatments about the questions asked in Wave 1, but were not informed of the response results.

Additionally, similar to the Control group, participants in the Norms treatment received the same reminders about the policy support questions, and participants in the Policy treatment received the reminders about the questions related to behavior and personal norms. Importantly, we did not repeat the information treatments in the third wave of the survey. All participants were simply reminded of the questions asked in the previous waves.

3.2.3 Post-treatment measures

In Waves 2 and 3, we collected two types of post-treatment outcome measures: immediate outcomes, measured in the second wave immediately after exposure to the reminders and treatments, and longer-term outcomes, measured three months later in the third wave. All participants provided the same outcome measures, regardless of treatment assignment. Additionally, in Wave 3, we presented participants with a set of questions aimed at testing their memory of the treatments. Specifically, we collected:

Personal norms (immediate). We asked participants to report the level they considered ethically appropriate for animal protein consumption ("From a sustainability viewpoint, how many meals per week containing meat, fish or seafood do you think it would be ethically appropriate to eat?"; scale ranging from 0 to 21),[2] home heating ("From a sustainability viewpoint, what do you think is the maximal temperature that it is ethically appropriate to set in a dwelling when it is less than 10 °C outside?"; scale ranging from 15 to 30), and use of public and soft mobility ("Please complete the following sentence: «In view of adopting a more sustainable lifestyle, it would be ethically appropriate for me to systematically switch to public transport or soft mobility as long as the additional time (in minutes, on top of the initial 30 minutes) does not exceed ... »; open-text box"). These questions were identical to those used in Wave 1 to elicit participants' personal norms. In both waves, the question on public and soft mobility was framed as a scenario in which all participants faced the same sacrifice, expressed as additional minutes compared to a 30-minute car trip. This approach is more effective for eliciting normative views than asking about the appropriate frequency of using public or soft mobility, as mobility patterns in Luxembourg are highly dependent on individual constraints (e.g., place of residence, workplace location, and access to reliable public transport). To ensure consistency in the directionality of our measures, we reverse-coded responses on animal protein consumption (21 - response) to obtain a measure of vegetarian meals per week. Similarly, we reverse-coded responses on home temperature (30 - response) to capture "degrees below the maximum." This approach ensures that higher values consistently reflect more sustainable norms.

Intended behaviors (immediate). We first reminded all participants of the levels of animal protein consumption, home heating reduction, and use of public and soft mobility that they had personally reported in Wave 1. We then asked them about their likelihood of changing these behaviors over the next three months, using a 4-point scale ranging from "0 - Completely certain of not changing" to "4 - Completely certain of changing." The only exception was home heating: since the survey was launched in early April and data collection lasted a month, with temperatures in Luxembourg typically becoming significantly warmer by the end of May, we asked participants about their intentions for the upcoming month instead. If participants selected any option above 0, they were then asked to indicate their intended behavior levels:

- Animal protein consumption: "How many meals containing meat, fish or seafood do you think you will eat per week in the upcoming 3 months?" (scale ranging from 0 to 21)

[2]We further informed participants that by "one meal containing meat, fish or seafood," we mean a meal containing at least 50 grams of it.

- Home heating: "What do you think will be the usual temperature of your dwelling when you are at home when it is less than 10 ℃, in the upcoming month? (in ℃)" (scale ranging from 15 to 30)
- Public transportation and soft mobility: "How frequently do you think you will use public transport (bus, train, tram) in the upcoming 3 months?" and "How frequently do you think you will use soft mobility (walk, cycle or use a (electric) scooter for a trip of 10 minutes or more) in the upcoming 3 months?" (5-point scale ranging from 1 - "Never" to 5 - "Daily")

We then created a single variable for each of the three intended behaviors (one for each behavior) by assigning participants the same behavioral level they reported in Wave 1 if they indicated being completely certain of not changing their behavior; otherwise, we used the new value reported in Wave 2. Additionally, we computed a single index of "public and soft mobility" by averaging the responses to both mobility questions. Finally, we reverse-coded the variables on meat consumption and home heating in the same way as we did for personal norms.

Actual behaviors (three months later). In Wave 3, we asked participants to report again their current levels of animal protein consumption ("Over the last 7 days, how many meals containing meat, fish or seafood did you eat?") and use of public transportation and soft mobility ("How frequently have you been using public transport (bus, train, tram) over the past two months?" and "How frequently have you been walking, cycling or using a (electric) scooter for a trip of 10 minutes or more, over the past two months?"). We used the same scales as those used for intended behaviors and reverse-coded responses on animal protein consumption. We also averaged responses to both mobility questions to create a single index of "public and soft mobility." We did not ask again about their home heating reduction, as Wave 3 was launched in July.

Policy support (immediate and three months later). In Wave 2, we provided all participants with detailed reminders of each of the six hypothetical policies presented in Wave 1 (and in the Policy treatment). We then asked participants again whether they would support each policy. In Wave 3, we gave a briefer reminder and asked the same question.

Memory about treatments (three months later). For participants assigned to the Norms and Policy treatment in Wave 2, we administered a memory task in Wave 3 to assess how well they recalled the information treatment. These memory tasks were presented after all outcomes of interest had been collected. For participants in the Norms treatment, we reminded them that in Wave 2 we had shared information about the sustainability-related behaviors and attitudes of other participants. We then presented three statements based on that information (one for each behavioral

domain) and asked them to indicate whether each statement was true or false. For those in the Policy treatment, we reminded them that in Wave 2 we had revealed which policies were supported by a majority in Wave 1. They were asked to indicate, for each of the six policies, whether it had been supported by a majority.

3.2.4 Addressing experimenter demand effects, social desirability bias and self-selection concerns

Similar to the protocol in Andre *et al.* (2024b), providing all groups with reminders about the Wave 1 questions during Wave 2 helps mitigate experimenter demand effects by ensuring that all participants, regardless of treatment assignment, are primed to think about similar issues. This approach also allows us to frame the information treatments as feedback on whether participants' previous responses were accurate, particularly since their social expectations were incentivized and eligible for a bonus based on correct guesses. As noted by Andre *et al.* (2024b) and Haaland, Roth and Wohlfart (2023), presenting information treatments as performance feedback tied to potential rewards is considered good practice to reduce experimenter demand effects. Notably, in Wave 3, all participants were exposed to the same information prior to the collection of our outcome measures. That is, we did not repeat the content of the information treatments provided in Wave 2, but simply reminded all participants of the questions asked in the previous waves and informed them that we would like to ask about these again. Thus, all participants were once again primed to think about the same issues, and any differences observed should reflect the lasting impact of belief updating that occurred in Wave 2. In line with this, there is consensus that conducting heterogeneity analyses by prior beliefs is a valuable strategy for identifying whether treatment effects are driven by belief updating (Andre *et al.* 2024b, Haaland, Roth and Wohlfart 2023).

Moreover, the same arguments presented in Chapter 2 of this dissertation apply here. According to De Quidt, Haushofer and Roth (2018), participants in online studies making incentivized and non-incentivized choices responded similarly to experimenter demand. In addition, non-representative online samples were no more prone to these effects than representative ones. The authors concluded that, in general, concerns about experimenter demand effects are limited in anonymized online surveys.

As noted by Reisinger (2022), anonymity in self-reports greatly decreases social desirability bias. Consistent with this, we assured participants that their responses would remain anonymous. Finally, as previously mentioned, a large-scale meta-analysis by Kormos and Gifford (2014) found

that self-reported pro-environmental behaviors are strongly correlated with objective behavioral measures.

To mitigate concerns about self-selection bias and the potential overrepresentation of participants with strong pro-environmental preferences, as stated in Chapter 2, we obscured the true purpose of the study when inviting participants to take part in Wave 1. Specifically, we informed them only that the survey would address their perceptions and behaviors related to societal challenges facing the Grand Duchy of Luxembourg. Although we could not control for sample attrition across waves, strong monetary incentives to participate in each wave help mitigate concerns that only individuals with strong social or pro-environmental preferences would remain in the study. Moreover, existing evidence suggests that the use of volunteer samples has a negligible impact on participants' social preferences and behavioral patterns (see, for example, Anderson *et al.* (2013); Falk, Meier and Zehnder (2013); Abeler and Nosenzo (2015)).

3.2.5 Sample

Table 3.1 shows the demographic characteristics of our sample. Income information was collected using intervals of €2,000. The median total net household monthly income falls within the €6,000-€8,000 interval. Based on this, we define "low income" as less than €6,000 and "high income" as more than €8,000. Since our sample is not representative of the Luxembourgish population, we applied weights to adjust for the population's age structure and gender composition in Table 3.1 and in all our subsequent analyses.

Table 3.1: Demographic distribution (weighted)

	Summary (N=912)
Low income (< €6,000)	0.429 (0.495)
High income (> €8,000)	0.275 (0.447)
Aged below 35	0.323 (0.468)
Aged above 65	0.154 (0.361)
Female	0.493 (0.500)
Higher education	0.654 (0.476)
Living in urban area	0.482 (0.500)

Mean proportion (SD in parentheses).

Table 3.2 shows the demographic characteristics of our sample by treatment.

Table 3.2: Demographic distribution, by treatment (weighted)

	Control (N=305)	Norms treatment (N=304)	Policy treatment (N=303)	Total (N=912)
Low income (< €6,000)	0.460 (0.499)	0.398 (0.490)	0.429 (0.496)	0.429 (0.495)
High income (> €8,000)	0.272 (0.446)	0.300 (0.459)	0.250 (0.434)	0.275 (0.447)
Aged below 35	0.298 (0.458)	0.336 (0.473)	0.335 (0.473)	0.323 (0.468)
Aged above 65	0.181 (0.386)	0.166 (0.373)	0.111 (0.315)	0.154 (0.361)
Female	0.510 (0.501)	0.509 (0.501)	0.458 (0.499)	0.493 (0.500)
Higher education	0.620 (0.486)	0.648 (0.478)	0.699 (0.459)	0.654 (0.476)
Living in urban area	0.508 (0.501)	0.417 (0.494)	0.528 (0.500)	0.482 (0.500)

Mean proportion (SD in parentheses).

3.3 Results

We now present the results of our information treatments. As described earlier, all our analyses were conducted using weights to adjust our sample to be representative of Luxembourg in terms of age and gender.[3] As shown in Table A3.1 in Appendix A3.1, participants' behaviors in Wave 1 (baseline) did not differ significantly between the treatment and control groups, indicating that the randomization of treatment allocation was successful. A post-hoc power analysis determined that, with our sample size, we had over 80% power to detect treatment effect sizes as small as Cohen's $f^2 = 0.01$ with $\alpha = 0.05$ in all of our main analyses (Cohen 1988).[4]

3.3.1 Model

As described in Section 3.2.3, we have four types of post-treatment outcomes of interest: (i) personal norms, measured immediately after the treatment in Wave 2; (ii) intended behavior in the upcoming months, measured immediately after the treatment in Wave 2; (iii) self-reported behavior, measured three months after the treatment in Wave 3; and (iv) policy support, measured in Waves 2 and 3.

For ease of interpretation, we standardized the outcome variables in (i)-(iii) to have a mean of zero and a standard deviation of one. Policy support is expressed as a binary variable taking values 0 and 1.

Following McKenzie (2012)'s recommendation, we estimate the following model:

[3]We did not consider additional demographic variables when creating our weights for two reasons: 1) data were not always readily available, and 2) each additional variable would inflate the variance of our weights and estimates, and we wanted to avoid excessive increases in variance.

[4]According to standard conventions, $f^2 = 0.02$ is defined as a small effect size (Cohen 1988).

$$y_i = \alpha + \beta_n T_i^n + \beta_p T_i^p + \gamma x_i + e_i \tag{3.1}$$

Our outcomes of interest are contained in y_i. T_i^n is a dummy variable taking value 1 for individuals in the Norms treatment group and 0 otherwise, and T_i^p is a dummy variable taking value 1 for individuals in the Policy treatment group and 0 otherwise. Finally, x_i is a vector of controls including age, gender, income, education, and the baseline values of our outcome variables measured in Wave 1. Outcomes (ii) and (iii) use the same baseline values as controls: self-reported behaviors in Wave 1. The error term is captured in e_i. The treatment effects are estimated here by β_n and β_p.

3.3.2 Main results

3.3.2.1 Personal norms

As shown in Table 3.3, correcting misperceptions about others' behaviors and norms influenced participants' personal norms regarding these behaviors.

Table 3.3: Personal norms for ethically appropriate behaviors (Wave 2)

	Vegetarian meals (PN) (1)	Lower home temperature (PN) (2)	Public mobility time (PN) (3)
Norms treatment	**0.265****	**0.232****	**-0.216***
Policy treatment	**0.207****	0.082	-0.064
Controls	Yes	Yes	Yes
Observations	912	912	912

* p<0.05, ** p<0.01, *** p<0.001.
Results from OLS regressions with weighted data. The dependent variable is a given personal norm.
Dependent variables are standardized: coefficients are expressed as SD.

Compared to the Control group, and controlling for baseline personal norms measured in Wave 1, participants in the Norms treatment increased the frequency of vegetarian consumption they considered ethically appropriate by 0.27 SD (p = 0.001) and the appropriate reduction in home temperature by 0.23 SD (p = 0.009). However, we also observed a negative effect: participants in the Norms treatment reduced the appropriate number of minutes they were willing to sacrifice to switch to public and soft mobility by 0.22 SD (p = 0.029). This negative effect is further explored in our heterogeneity analyses in Section 3.3.3.

Additionally, we found positive spillover effects: participants exposed to the Policy treatment also increased their appropriate vegetarian consumption frequency by 0.21 SD (p = 0.008).

3.3.2.2 Behaviors

As shown in Table 3.4, correcting misperceptions about other participants' behaviors and norms was effective in influencing participants' behaviors, both intended and actual, with the exception of the use of public and soft mobility.

Table 3.4: Intended and actual behaviors

	Vegetarian meals (1)	Lower home temperature (2)	Public & soft mobility (3)
Panel A: Intended behaviors			
Norms treatment	**0.112***	**0.197****	-0.035
Policy treatment	-0.019	0.092	-0.014
Panel B: Actual behaviors			
Norms treatment	**0.145***		0.034
Policy treatment	0.064		0.060
Controls	Yes	Yes	Yes
Observations	912	912	912

* p<0.05, ** p<0.01, *** p<0.001.

Results from OLS regressions with weighted data. The dependent variable in Panel A is a given intended behavior measured immediately in Wave 2, and in Panel B is a given self-reported behavior measured three months later in Wave 3. Dependent variables are standardized: coefficients are expressed as SD.

As observed in Panel A of Table 3.4, compared to the Control group, participants in the Norms treatment intended to increase their weekly vegetarian consumption frequency by 0.11 SD (p = 0.038, Column 1) and their home heating reduction efforts by 0.20 SD (p = 0.005, Column 2). There were no differences in the intended use of public and soft mobility between the Control and Norms groups (Column 3), nor were there spillover effects from the Policy treatment.

Three months after the initial exposure to the treatments, participants in the Norms group upheld their intentions and reported a significantly higher weekly vegetarian consumption frequency compared to the Control group (0.15 SD, p = 0.039, Column 1 in Panel B). Again, consistent with their intentions, no differences were found between the Norms and Control groups in public and soft mobility, and no spillover effects from the Policy treatment were observed.

3.3.2.3 Policy support

As confirmed by Table A3.2 in Appendix A3.2, the treatments did not affect the likelihood of supporting the policies that were revealed in the Policy treatment to be supported by a minority in Wave 1, either when asked immediately after the treatments or three months later. Table 3.5, however, shows that correcting misperceptions about others' policy support increased participants'

likelihood of supporting policies that were revealed to be supported by a majority in Wave 1, particularly in the domains of vegetarian consumption and home heating.

Table 3.5: Policy support: Policies supported by majority

	Regulation on red meat (1)	Fossil fuel rationing (2)	Rental tax on poor insulation (3)	Ban on cars in city center (4)
Panel A: Immediate				
Norms treatment	0.061	0.012	-0.008	-0.082
Policy treatment	**0.134****	**0.092***	0.066	0.022
Panel B: After 3 months				
Norms treatment	**0.100***	0.060	-0.002	-0.002
Policy treatment	**0.097***	0.082	-0.020	-0.003
Controls	Yes	Yes	Yes	Yes
Observations	912	912	912	912

* p<0.05, ** p<0.01, *** p<0.001.

Results from OLS regressions with weighted data. The dependent variable in Panel A is support for a given policy measured immediately in Wave 2, and in Panel B is support for a given policy measured three months later in Wave 3. Dependent variables are expressed as binary variables with values 0 and 1. Coefficients, hence, can be interpreted as probabilities.

Participants in the Policy treatment group were immediately 13 percentage points more likely to support a regulation on red meat (p = 0.001, Column 1 in Panel A) and 9 percentage points more likely to support a fossil fuel rationing (p = 0.041, Column 2 in Panel A) compared to the Control group. No differences were found between the Policy and Control groups for the other two hypothetical policies. No spillover effects from the Norms treatment were found either.

When asked again three months later in Wave 3, the positive treatment effect on the likelihood of supporting a regulation on red meat persisted: participants in the Policy treatment group were still 10 percentage points more likely to support the policy compared to the Control group (p = 0.018, Column 1 in Panel B). However, the increase in the likelihood of supporting fossil fuel rationing was no longer significant and, consistent with the immediate effects, neither was the likelihood of supporting the remaining two policies. Interestingly, although no spillover effects were observed in Wave 2, by Wave 3, participants in the Norms treatment group were also 10 percentage points more likely to support a regulation on red meat compared to the Control group (p = 0.025, Column 1 in Panel B).

3.3.3 Treatment heterogeneity

We examined whether our treatments had heterogeneous effects based on prior beliefs by replicating our previous analyses separately for participants who held pessimistic beliefs about a given personal norm, behavior, or policy support in Wave 1, and for those who did not. This approach is similar to that followed by Andre *et al.* (2024b).

We also examined whether our treatments had heterogeneous effects based on prior behavior by replicating our previous analyses separately for participants who reported a sustainable behavior level above the average and those who reported a level below the average in Wave 1. For policy support, we conducted the analyses separately for participants who supported a given policy and those who did not. In this case, we could not use baseline policy support as a control variable in our regression analyses, as it served as a source of heterogeneity.

The main advantage of conducting our analyses separately based on prior measures is that we can observe not only whether the treatments were equally likely to be effective across all subgroups of participants, but also whether the treatments backfired for a given subgroup.

3.3.3.1 Personal norms

As shown in Table 3.6, the Norms treatment was particularly effective at increasing the appropriate vegetarian consumption frequency among participants who underestimated others' personal norms on this dimension in Wave 1 (0.38 SD, $p < 0.001$, Column 1). The same pattern applied to home heating reduction efforts (0.45 SD, $p = 0.001$, Column 3). In both cases, the treatment had no effect among those who did not hold pessimistic prior beliefs. Conversely, the Norms treatment decreased the appropriate time sacrificed to switch to public and soft mobility only among the non-pessimistic (-0.36 SD, $p = 0.036$, Column 6), while those with pessimistic prior beliefs were unaffected.

We also find evidence of a positive spillover effect from the Policy treatment on appropriate vegetarian consumption frequency among participants who did not hold pessimistic prior beliefs (0.28 SD, $p = 0.013$).

As shown in Table 3.7, the Norms treatment was effective in increasing the appropriate vegetarian consumption frequency for both sustainable (0.34 SD, $p = 0.009$, Column 1) and unsustainable participants (0.22 SD, $p = 0.022$, Column 2). It increased appropriate home heating efforts only among the unsustainable participants (0.44 SD, $p = 0.006$, Column 3). Mirroring the results of our heterogeneity analysis by prior beliefs, the Norms treatment backfired in the appropriate time sacrificed for public mobility, but only among the sustainable participants (-0.33 SD, $p = 0.012$,

Table 3.6: Personal norms for ethically appropriate behaviors (heterogeneity by prior beliefs)

	Vegetarian meals (PN)		Lower home temperature (PN)		Public mobility time (PN)	
	Pessimistic	Non-pessimistic	Pessimistic	Non-pessimistic	Pessimistic	Non-pessimistic
	(1)	(2)	(3)	(4)	(5)	(6)
Norms treatment	**0.379*****	0.146	**0.449****	-0.070	-0.101	**-0.364***
Policy treatment	0.144	**0.285***	0.148	0.001	0.029	-0.177
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	476	436	492	420	560	352

* p<0.05, ** p<0.01, *** p<0.001.

Results from OLS regressions with weighted data. The dependent variable is a given personal norm. Data is subset by prior beliefs. Dependent variables are standardized: coefficients are expressed as SD.

Column 6).

Table 3.7: Personal norms for ethically appropriate behaviors (heterogeneity by prior behavior)

	Vegetarian meals (PN)		Lower home temperature (PN)		Public mobility time (PN)	
	< mean	> mean	< mean	> mean	< mean	> mean
	(1)	(2)	(3)	(4)	(5)	(6)
Norms treatment	**0.340****	**0.219***	**0.437****	0.090	-0.061	**-0.328***
Policy treatment	0.219	0.201	0.152	0.029	-0.082	-0.073
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	392	520	349	563	361	551

* p<0.05, ** p<0.01, *** p<0.001.

Results from OLS regressions with weighted data. The dependent variable is a given personal norm. Data is subset by prior behavior. Dependent variables are standardized: coefficients are expressed as SD.

3.3.3.2 Behaviors

While the Norms treatment did not affect intentions to increase vegetarian consumption differently between pessimistic and non-pessimistic participants, it increased intentions to reduce home heating only among the pessimistic (0.19 SD, p = 0.038; Column 3 in Table 3.8, Panel A). Similarly, it increased actual vegetarian consumption in Wave 3 only among the pessimistic (0.24 SD, p = 0.009; Column 1 in Panel B).

We also detected positive spillover effects from the Policy treatment, again among the non-pessimistic, for vegetarian consumption in Wave 3 (0.22 SD, p = 0.02; Column 2 in Panel B).

Replicating our results with prior beliefs, the Norms treatment increased intentions to reduce home heating (0.30 SD, p = 0.011; Column 3 in Table 3.9, Panel A) and actual vegetarian consumption in Wave 3 (0.26 SD, p = 0.035; Column 1 in Panel B) only among the unsustainable.

Table 3.8: Intended and actual behaviors (heterogeneity by prior beliefs)

	Vegetarian meals		Lower home temperature		Public & soft mobility	
	Pessimistic	Non-pessimistic	Pessimistic	Non-pessimistic	Pessimistic	Non-pessimistic
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A:						
Intended behaviors						
Norms treatment	0.098	0.118	**0.187***	0.197	0.097	-0.119
Policy treatment	-0.021	-0.006	0.144	-0.001	0.013	-0.035
Panel B:						
Actual behaviors						
Norms treatment	**0.237****	0.040			0.025	0.054
Policy treatment	-0.025	**0.223***			0.126	0.027
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	553	359	541	371	393	519

* p<0.05, ** p<0.01, *** p<0.001.

Results from OLS regressions with weighted data. The dependent variable in Panel A is a given intended behavior measured immediately in Wave 2, and in Panel B is a given self-reported behavior measured three months later in Wave 3. Data is subset by prior beliefs. Dependent variables are standardized: coefficients are expressed as SD.

3.3.3.3 Policy support

Table 3.10 shows that the Policy treatment increased the likelihood of supporting a regulation on red meat both immediately (15 pp, $p < 0.001$; Column 1 in Panel A) and three months later (10 pp, $p = 0.016$; Column 1 in Panel B) only among the pessimistic. It also increased the likelihood of supporting fossil fuel rationing in Wave 3 again only among the pessimistic (11 pp, $p = 0.029$; Column 3 in Panel B).

We also observe positive spillover effects from the Norms treatment: although it had no immediate effect, three months later it increased support for red meat regulation among the pessimistic (10 pp, $p = 0.028$; Column 1 in Panel B).

Finally, heterogeneity analyses based on prior support show similar patterns to those observed with prior beliefs. The Policy treatment increased the likelihood of supporting a regulation on red meat only among those who previously did not support the policy, both immediately (23 pp, $p = 0.002$; Column 1 in Table 3.11, Panel A) and three months later (15 pp, $p = 0.025$; Column 1 in Panel B).

We also observed positive spillover effects from the Norms treatment: in Wave 3, it increased the likelihood to support a red meat regulation among prior detractors (20 pp, $p = 0.008$; Column 1 in Panel B) and increased the likelihood to support a fossil fuel rationing among prior supporters (13 pp, $p = 0.038$; Column 4 in Panel B).

Table 3.9: Intended and actual behaviors (heterogeneity by prior behavior)

	Vegetarian meals		Lower home temperature		Public & soft mobility	
	< mean	> mean	< mean	> mean	< mean	> mean
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Intended behaviors						
Norms treatment	0.176	0.080	**0.304***	0.116	0.038	-0.060
Policy treatment	0.039	-0.052	0.033	0.102	-0.015	-0.007
Panel B: Actual behaviors						
Norms treatment	**0.264***	0.093			-0.064	0.073
Policy treatment	0.007	0.116			0.048	0.049
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	392	520	349	563	361	551

* p<0.05, ** p<0.01, *** p<0.001.

Results from OLS regressions with weighted data. The dependent variable in Panel A is a given intended behavior measured immediately in Wave 2, and in Panel B is a given self-reported behavior measured three months later in Wave 3. Data is subset by prior behavior. Dependent variables are standardized: coefficients are expressed as SD.

3.3.4 Treatment memory

Finally, we examined whether participants were more likely to increase their sustainable behaviors and policy support in Wave 3 if they better remembered the content of the treatments. Overall, most participants demonstrated poor recall of the treatment content (see Appendix A3.3 for details on their performance in the memory task). However, participants were more likely to correctly identify the Norms treatment information about animal protein consumption (37%) than about public and soft mobility (19%; Wilcoxon signed-rank test: $p < 0.001$). This prompted us to explore whether memory of a treatment could help explain the asymmetric treatment effects observed between vegetarian consumption and the use of public and soft mobility.

As shown in Table A3.3 in Appendix A3.4, correctly answering the memory question about animal protein consumption is indeed associated with an increase in self-reported vegetarian consumption frequency of 0.27 SD ($p = 0.006$). However, we did not find a significant association between correctly recalling the information about public and soft mobility and self-reported use of such mobility. Thus, while treatment memory appears to amplify the Norms treatment's effects on vegetarian consumption, it does not fully account for its differential effectiveness across behaviors.

On the other hand, as shown in Table A3.4 in Appendix A3.4, correctly recalling that a given policy was supported by the majority increased participants' likelihood of supporting that policy by 21 to 30 percentage points (all $p \leq 0.001$). Consistent with our expectations that treatment effects operate through positive belief updating, correctly remembering that the highway toll policy was supported by a minority in Wave 1 is associated with a 33 percentage point decrease in support for that policy ($p < 0.001$). No such association was found for the VAT on meat policy, however.

Table 3.10: Policy support: Policies supported by majority (heterogeneity by prior beliefs)

	Regulation on red meat		Fossil fuel rationing		Rental tax on poor insulation		Ban on cars in city center	
	Pess.	Non-pess.	Pess.	Non-pess.	Pess.	Non-pess.	Pess.	Non-pess.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Immediate								
Norms treatment	0.074	-0.022	0.022	-0.002	0.006	-0.055	-0.067	-0.118
Policy treatment	**0.146*****	-0.074	0.092	0.132	0.078	0.009	0.016	0.071
Panel B: After 3 months								
Norms treatment	**0.100***	0.190	0.048	0.134	0.002	-0.031	-0.027	0.071
Policy treatment	**0.103***	-0.013	**0.111***	0.037	-0.054	0.064	-0.014	0.015
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	827	85	736	176	679	233	717	195

* p<0.05, ** p<0.01, *** p<0.001.

Results from OLS regressions with weighted data. Data is subset by prior beliefs. The dependent variable in Panel A is support for a given policy measured immediately in Wave 2, and in Panel B is support for a given policy measured three months later in Wave 3. Dependent variables are expressed as binary variables with values 0 and 1. Coefficients, hence, can be interpreted as probabilities.

Despite this positive association, participants were no more likely to remember that a regulation on red meat, the policy for which we observed the strongest increase in support in response to the Policy treatment, was supported by a majority, compared to the other policies. In fact, the share of participants who correctly remembered that this policy was supported by a majority (28%) was lower than the share who correctly remembered that the other three policies were supported by a majority (31%, 36%, and 30%). Hence, once again, while treatment memory seems to amplify the positive effects of the Policy treatment on policy support, it cannot explain the variation in treatment effectiveness across policies.

Table 3.11: Policy support: Policies supported by majority (heterogeneity by prior support)

	Regulation on red meat		Fossil fuel rationing		Rental tax on poor insulation		Ban on cars in city center	
	Detract	Support	Detract	Support	Detract	Support	Detract	Support
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Immediate								
Norms treatment	0.093	0.036	0.027	0.002	0.021	-0.027	-0.101	-0.052
Policy treatment	**0.226****	0.070	0.078	0.097	0.127	0.028	0.007	0.053
Panel B: After 3 months								
Norms treatment	**0.196****	0.046	-0.015	**0.134***	0.128	-0.069	0.008	-0.007
Policy treatment	**0.150***	0.073	0.105	0.066	0.078	-0.063	0.013	-0.028
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	343	569	448	464	317	595	458	454

* p<0.05, ** p<0.01, *** p<0.001.

Results from OLS regressions with weighted data. Data is subset by prior support. The dependent variable in Panel A is support for a given policy measured immediately in Wave 2, and in Panel B is support for a given policy measured three months later in Wave 3. Dependent variables are expressed as binary variables with values 0 and 1. Coefficients, hence, can be interpreted as probabilities.

3.3.5 Discussion

The targeted treatments immediately increased participants' sustainable personal norms, intended behaviors in the upcoming months, and support for regulatory policies that had majority support in Wave 1, but only in the domains of vegetarian consumption and home heating. For vegetarian consumption, the effects on policy support persisted for at least three months, and participants followed through on their intentions by reporting an actual increase in the frequency of vegetarian consumption in Wave 3. In contrast, participants did not respond to the treatments in the domain of public and soft mobility in terms of behavior or policy support and showed a negative response in terms of sustainable personal norms. As expected, the treatments did not increase support for policies that had minority support in Wave 1. Overall, taxation policies appear to be less popular than regulatory policies and less responsive to our treatments.

Our results align with our expectation that the effectiveness of the treatments stems from positive belief updating and upward norm adjustment. Indeed, our heterogeneity analyses based on prior beliefs and behavior show that our targeted treatments were more likely to increase personal norms, behaviors, and policy support among participants who had previously underestimated the sustainable norm, reported below-average sustainable behaviors, or did not support the policy. In contrast, the targeted treatments generally had no positive effect on participants who either did not update their beliefs or updated them negatively (i.e., those who had previously overestimated the sustainable norm). Similarly, the targeted treatments rarely had a positive effect on participants who already behaved sustainably in Wave 1. Our heterogeneity analyses also provide more insight into the backfiring effect of the Norms treatment on personal norms related to public and soft mobility. The treatment decreased personal norms only among participants who did not hold prior pessimistic beliefs (and thus possibly updated their beliefs negatively) and who previously reported above-average behavior in this domain.

However, we did not observe this backfiring effect on personal norms in the other two behavioral domains. Importantly, we also did not observe any backfiring for behaviors or policy support. That is, non-pessimistic and sustainable participants did not reduce their sustainable behaviors to conform to the perceived norm (consistent with the results of Andre *et al.* (2024b)), nor did they decrease their likelihood of supporting a policy. In fact, our participants did not decrease their likelihood to support a policy even when the policy was revealed to be supported by a minority.

Further evidence that updating participants' social expectations was a necessary condition for the effectiveness of our targeted treatments, rather than pessimistic and unsustainable participants

being more sensitive to treatments overall, comes from our observation that whenever we detected positive spillover effects from the non-targeted treatment, this pattern no longer applied. That is, non-targeted treatments were not more likely to increase behaviors, personal norms, or policy support among pessimistic participants or those who behaved unsustainably. This result aligns with the fact that non-targeted treatments did not correct participants' beliefs about the specific behavior, norm, or policy, nor did they provide information about a norm for participants to conform to regarding the specific outcome measure.

Our results also provide evidence that interventions promoting sustainable behavior and those promoting policy support do not necessarily act as substitutes in promoting behavioral change in a community. Specifically, we found no evidence of negative spillover effects from the non-targeted treatment in any of our outcomes of interest. Instead, we observed either no spillover effects or positive ones.

Finally, it is noteworthy that vegetarian consumption was the behavioral domain in which participants were most open to changing their behaviors, personal norms, and policy support, whereas public and soft mobility was the domain that faced the greatest resistance to positive change across all outcome measures. Through our analyses, we rule out three possible explanations for why the treatments were ineffective in this particular domain:

1. The lack of effectiveness cannot be attributed to the prevalence of positive belief updating, as treatments were no more effective for participants who held prior pessimistic beliefs about public and soft mobility.

2. It cannot be explained by the availability of upward behavior adjustment, as treatments were not more effective for the most unsustainable participants in this domain, either in terms of behavior or policy support.

3. It also cannot be explained by participants' likelihood of remembering the treatment content (and thus the norm).

Having ruled out these potential explanations, we propose a remaining plausible interpretation. Correcting misperceptions may be more effective in changing behaviors within domains that are less costly to adopt, such as increasing vegetarian consumption. In contrast, behaviors that are more difficult to change and involve greater personal constraints, such as mobility patterns, may require more substantial structural support from institutions. Mobility patterns may depend on access to reliable alternatives to cars or specific commuting needs. Indeed, despite the Luxembourgish

government's implementation of strong economic incentives to reduce car use, such as making all public transportation within the country completely free of charge since 2020 (Rose 2023), car usage in Luxembourg remains among the highest in the EU (Heindrichs 2024).

3.4 Conclusion

This study demonstrates the potential of norm-based interventions, and more specifically, interventions that correct misperceived social norms, to promote structural behavioral and policy change in support of a transition toward a low-carbon economy. Using a three-wave longitudinal design over nine months in Luxembourg, we show that participants systematically underestimated the sustainability of their peers. Providing corrective information through targeted treatments led to meaningful and lasting changes in personal norms, behaviors, and policy preferences, particularly in domains where behavioral change is relatively low-cost, such as vegetarian consumption.

Our findings suggest that the success of these interventions depends on positive belief updating and the potential for upward behavioral adjustment, conditions that are more likely to occur in contexts characterized by pluralistic ignorance (Prentice and Miller 1996) or false consensus effects (Ross, Greene and House 1977). Notably, we find no evidence of behavioral and policy support backfiring effects, even among individuals who may have updated their beliefs negatively, were already behaving sustainably, or learned that the policy had only minority support. However, we did observe a backfiring effect on personal norms related to public and soft mobility among participants who did not hold prior pessimistic beliefs and reported above-average sustainable behavior. These findings call for caution: while correcting misperceptions can serve as an effective policy tool to promote behavioral change, it is important to examine potential unintended effects on non-behavioral measures, such as personal norms.

Furthermore, we find that interventions promoting sustainable behaviors and those encouraging policy support are not substitutive. This finding supports the potential of integrated strategies and contributes to the broader discussion on spillover effects and crowding out between sustainable behaviors and sustainable policy preferences (Truelove *et al.* 2014, Maki *et al.* 2019, Sparkman, Attari and Weber 2021, Werfel 2017). At the same time, the domain-specific nature of our positive treatment effects highlights the need for additional structural support in sectors where individual behavioral change faces greater barriers, such as mobility, which is often constrained by infrastructure or limited access to viable alternatives.

Taken together, our results offer practical insights for the design of norm-based interventions

and add to the growing body of evidence on the role of social expectations in shaping sustainable preferences and decision-making (Andre *et al.* 2024a,b, Ferraro, Miranda and Price 2011, Geiger and Swim 2016, Vesely and Klöckner 2018). These insights are valuable both for policymakers seeking to promote behavioral change and support the transition toward a low-carbon society, and for researchers studying social norms and their influence on individual and collective preferences.

Future research could continue to explore the boundaries and limitations of norm-based interventions and deepen our understanding of their mechanisms, for example, by examining spillover effects across different sustainable behaviors. Although our study investigates longer-term impacts, it covers only a three-month period; future studies could assess the durability of these effects over more extended time frames. Another promising direction is to test interventions that correct misperceived social norms using more robust behavioral measures, such as administrative data, rather than relying on self-reports, to better evaluate the robustness and replicability of our findings.

References

Abeler, Johannes, and Daniele Nosenzo. 2015. "Self-selection into laboratory experiments: pro-social motives versus monetary incentives." Experimental Economics, 18: 195–214.

Anderson, Jon, Stephen V Burks, Jeffrey Carpenter, Lorenz Götte, Karsten Maurer, Daniele Nosenzo, Ruth Potter, Kim Rocha, and Aldo Rustichini. 2013. "Self-selection and variations in the laboratory measurement of other-regarding preferences across subject pools: evidence from one college student and two adult samples." Experimental Economics, 16: 170–189.

Andre, Peter, Teodora Boneva, Felix Chopra, and Armin Falk. 2024a. "Globally representative evidence on the actual and perceived support for climate action." Nature Climate Change, 14(3): 253–259.

Andre, Peter, Teodora Boneva, Felix Chopra, and Armin Falk. 2024b. "Misperceived social norms and willingness to act against climate change." Review of Economics and Statistics, 1–46.

Bašić, Zvonimir, and Eugenio Verrina. 2024. "Personal norms—and not only social norms—shape economic behavior." Journal of Public Economics, 239: 105255.

Bicchieri, Cristina. 2005. The grammar of society: The nature and dynamics of social norms. Cambridge University Press.

Bicchieri, Cristina. 2016. Norms in the wild: How to diagnose, measure, and change social norms. Oxford University Press.

Bursztyn, Leonardo, Alessandra L González, and David Yanagizawa-Drott. 2020. "Misperceived social norms: Women working outside the home in Saudi Arabia." American economic review, 110(10): 2997–3029.

Cialdini, Robert B, and Ryan P Jacobson. 2021. "Influences of social norms on climate change-related behaviors." Current Opinion in Behavioral Sciences, 42: 1–8.

Cohen, Jacob. 1988. Statistical power analysis for the behavioral sciences. routledge.

De Groot, Judith IM, Wokje Abrahamse, and Kayleigh Jones. 2013. "Persuasive normative messages: The influence of injunctive and personal norms on using free plastic bags." Sustainability, 5(5): 1829–1844.

De Quidt, Jonathan, Johannes Haushofer, and Christopher Roth. 2018. "Measuring and bounding experimenter demand." American Economic Review, 108(11): 3266–3302.

Deutchman, Paul, Gordon Kraft-Todd, Liane Young, and Katherine McAuliffe. 2024. "People update their injunctive norm and moral beliefs after receiving descriptive norm information." Journal of Personality and Social Psychology.

Druckman, Angela, and Tim Jackson. 2016. "Understanding households as drivers of carbon emissions." Taking stock of industrial ecology, 181–203.

Dubois, Ghislain, Benjamin Sovacool, Carlo Aall, Maria Nilsson, Carine Barbier, Alina Herrmann, Sébastien Bruyère, Camilla Andersson, Bore Skold, Franck Nadaud, *et al.* 2019. "It starts at home? Climate policies targeting household consumption and behavioral decisions are key to low-carbon futures." Energy Research & Social Science, 52: 144–158.

Falk, Armin, Stephan Meier, and Christian Zehnder. 2013. "Do lab experiments misrepresent social preferences? The case of self-selected student samples." Journal of the European Economic Association, 11(4): 839–852.

Ferraro, Paul J, Juan Jose Miranda, and Michael K Price. 2011. "The persistence of treatment effects with norm-based policy instruments: evidence from a randomized environmental policy experiment." American Economic Review, 101(3): 318–322.

Geiger, Nathaniel, and Janet K Swim. 2016. "Climate of silence: Pluralistic ignorance as a barrier to climate change discussion." Journal of Environmental Psychology, 47: 79–90.

Goldstein, Noah J, Vladas Griskevicius, and Robert B Cialdini. 2007. "Invoking social norms: A social psychology perspective on improving hotels' linen-reuse programs." Cornell Hotel and Restaurant Administration Quarterly, 48(2): 145–150.

Gravert, Christina, and Linus Olsson Collentine. 2021. "When nudges aren't enough: Norms, incentives and habit formation in public transport usage." Journal of Economic Behavior & Organization, 190: 1–14.

Griesoph, Amelie, Stefan Hoffmann, Christine Merk, Katrin Rehdanz, and Ulrich Schmidt. 2021. "Guess What. . . ?—How Guessed Norms Nudge Climate-Friendly Food Choices in Real-Life Settings." Sustainability, 13(15): 8669.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart. 2023. "Designing information provision experiments." Journal of economic literature, 61(1): 3–40.

Hallsworth, Michael, John A List, Robert D Metcalfe, and Ivo Vlaev. 2017. "The behavioralist as tax collector: Using natural field experiments to enhance tax compliance." Journal of public economics, 148: 14–31.

Heindrichs, Tracy. 2024. "Luxembourg car usage remains among EU's highest." Luxembourg Times. `https://www.luxtimes.lu/europeanunion/luxembourg-car-usage-remains-among-eu-s-highest/20106555.html`.

Hertwich, Edgar G, and Glen P Peters. 2009. "Carbon footprint of nations: a global, trade-linked analysis." Environmental science & technology, 43(16): 6414–6420.

Hitaj, Claudia, Elorri Igos, and Thomas Gibon. 2022. "Towards decabornisation: Understanding and reducing our carbon footprint in Luxembourg." Rapports du LIST.

Khalfan, Ashfaq, Astrid Nilsson Lewis, Carlos Aguilar, Jacqueline Persson, Max Lawson, Nafkote Dabi, Safa Jayoussi, and Sunil Acharya. 2023. "Climate Equality: A planet for the 99%."

Kormos, Christine, and Robert Gifford. 2014. "The validity of self-report measures of proenvironmental behavior: A meta-analytic review." Journal of Environmental Psychology, 40: 359–371.

Kormos, Christine, Robert Gifford, and Erinn Brown. 2015. "The influence of descriptive social norm information on sustainable transportation behavior: A field experiment." Environment and Behavior, 47(5): 479–501.

Lindström, Björn, Simon Jangard, Ida Selbing, and Andreas Olsson. 2018. "The role of a "common is moral" heuristic in the stability and change of moral norms." Journal of Experimental Psychology: General, 147(2): 228.

Maki, Alexander, Amanda R Carrico, Kaitlin T Raimi, Heather Barnes Truelove, Brandon Araujo, and Kam Leung Yeung. 2019. "Meta-analysis of pro-environmental behaviour spillover." Nature Sustainability, 2(4): 307–315.

McKenzie, David. 2012. "Beyond baseline and follow-up: The case for more T in experiments." Journal of development Economics, 99(2): 210–221.

Mernyk, Joseph S, Sophia L Pink, James N Druckman, and Robb Willer. 2022. "Correcting inaccurate metaperceptions reduces Americans' support for partisan violence." Proceedings of the National Academy of Sciences, 119(16): e2116851119.

Nolan, Jessica M, P Wesley Schultz, Robert B Cialdini, Noah J Goldstein, and Vladas Griskevicius. 2008. "Normative social influence is underdetected." Personality and social psychology bulletin, 34(7): 913–923.

Prentice, Deborah A, and Dale T Miller. 1993. "Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm." Journal of personality and social psychology, 64(2): 243.

Prentice, Deborah A, and Dale T Miller. 1996. "Pluralistic ignorance and the perpetuation of social norms by unwitting actors." In Advances in experimental social psychology. Vol. 28, 161–209. Elsevier.

Reese, Gerhard, Kristina Loew, and Georges Steffgen. 2014. "A towel less: Social norms enhance pro-environmental behavior in hotels." The Journal of Social Psychology, 154(2): 97–100.

Reisinger, James. 2022. "Subjective well-being and social desirability." Journal of public economics, 214: 104745.

Richter, Isabel, John Thøgersen, and Christian A Klöckner. 2018. "A social norms intervention going wrong: Boomerang effects from descriptive norms information." Sustainability, 10(8): 2848.

Rose, Steve. 2023. "All aboard! Can Luxembourg's free public transport help save the world?" The Guardian. `https://www.theguardian.com/world/2023/sep/20/all-aboard-can-luxembourgs-free-public-transport-help-save-the-world`.

Ross, Lee, David Greene, and Pamela House. 1977. "The "false consensus effect": An egocentric bias in social perception and attribution processes." Journal of experimental social psychology, 13(3): 279–301.

Saracevic, Selma, and Bodo B Schlegelmilch. 2021. "The impact of social norms on pro-environmental behavior: a systematic literature review of the role of culture and self-construal." Sustainability, 13(9): 5156.

Schultz, P Wesley. 1999. "Changing behavior with normative feedback interventions: A field experiment on curbside recycling." Basic and applied social psychology, 21(1): 25–36.

Schultz, Wesley P, Azar M Khazian, and Adam C Zaleski. 2008. "Using normative social influence to promote conservation among hotel guests." Social influence, 3(1): 4–23.

Sokoloski, Rebecca, Ezra M Markowitz, and David Bidwell. 2018. "Public estimates of support for offshore wind energy: False consensus, pluralistic ignorance, and partisan effects." Energy Policy, 112: 45–55.

Sparkman, Gregg, Bobbie NJ Macdonald, Krystal D Caldwell, Brian Kateman, and Gregory D Boese. 2021. "Cut back or give it up? The effectiveness of reduce and eliminate appeals and dynamic norm messaging to curb meat consumption." Journal of Environmental Psychology, 75: 101592.

Sparkman, Gregg, Nathan Geiger, and Elke U Weber. 2022. "Americans experience a false social reality by underestimating popular climate policy support by nearly half." Nature communications, 13(1): 4779.

Sparkman, Gregg, Shahzeen Z Attari, and Elke U Weber. 2021. "Moderating spillover: Focusing on personal sustainable behavior rarely hinders and can boost climate policy support." Energy Research & Social Science, 78: 102150.

Swim, Janet K, and Nathaniel Geiger. 2021. "Policy attributes, perceived impacts, and climate change policy preferences." Journal of Environmental Psychology, 77: 101673.

Truelove, Heather Barnes, Amanda R Carrico, Elke U Weber, Kaitlin Toner Raimi, and Michael P Vandenbergh. 2014. "Positive and negative spillover of pro-environmental behavior: an integrative review and theoretical framework." Global Environmental Change, 29: 127–138.

Verheyden, Bertrand, Michel Tenikue, Philippe Van Kerm, Ángela Jiang-Wang, Francesco Fallucchi, and David Cristelo. 2024. "Driving Behavioral Change for an Economic and Social Transition towards more Resilience and Sustainability in Luxembourg: SOC2050." Rapports du LISER.

Vesely, Stepan, and Christian A Klöckner. 2018. "Global social norms and environmental behavior." Environment and Behavior, 50(3): 247–272.

Werfel, Seth H. 2017. "Household behaviour crowds out support for climate change policy when sufficient progress is perceived." Nature Climate Change, 7(7): 512–515.

Appendix

A3.1 Baseline behaviors, by treatment

Table A3.1: Behaviors in wave 1, by treatment

	Vegetarian meals (1)	Lower home temperature (2)	Public & soft mobility (3)
Norms treatment	0.147	0.076	-0.049
Policy treatment	0.135	0.094	0.041
Controls	Yes	Yes	Yes
Observations	912	912	912

* p<0.05, ** p<0.01, *** p<0.001. Outcome variables standardized with a mean of 0 and standard deviation of 1.

A3.2 Treatment effect on support for policies that were revealed to be supported by a minority

Table A3.2: Policy support: Policies supported by minority

	VAT on meat (1)	Toll on highways (2)
Panel A: Immediate		
Norms treatment	-0.004	-0.052
Policy treatment	0.032	0.023
Panel B: After 3 months		
Norms treatment	-0.011	-0.016
Policy treatment	0.034	-0.014
Controls	Yes	Yes
Observations	912	912

* p<0.05, ** p<0.01, *** p<0.001.

Results from OLS regressions with weighted data. The dependent variable in Panel A is support for a given policy measured immediately in Wave 2, and in Panel B is support for a given policy measured three months later in Wave 3. Dependent variables are expressed as binary variables with values 0 and 1. Coefficients, hence, can be interpreted as probabilities.

A3.3 Memory of treatments

Participants assigned to the Norms treatment in Wave 2 were asked in Wave 3 to indicate whether three statements, related to the information they received in Wave 2, were true or false. Figure A3.1 presents a histogram showing the average proportion of correct responses. Only 12.8% of participants answered at least two statements correctly, while 35.5% answered none correctly.

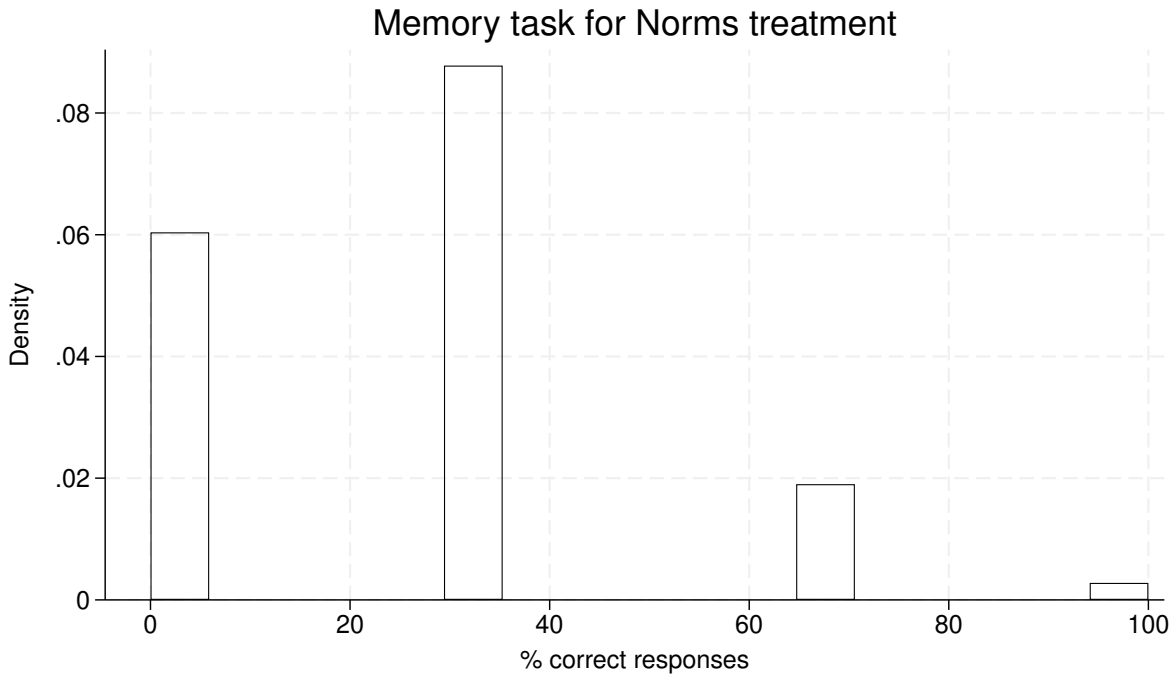


Figure A3.1: Percentage of correct responses in memory task

Participants assigned to the Policy treatment in Wave 2 were asked in Wave 3 to identify which four out of six policies had majority support in Wave 1. Only 3 out of 303 participants (1%) correctly identified all four policies. Additionally, 38.61% of participants selected "I really don't remember."

A3.4 Does remembering the content of the treatments predict behavior and policy support?

Table A3.3: Does remembering the norm predict behavior?

	Vegetarian meals (1)	Public & soft mobility (2)
Correct recall	**0.274****	0.084
Controls	Yes	Yes
Observations	304	304

* p<0.05, ** p<0.01, *** p<0.001.

Results from OLS regressions with weighted data. The dependent variable is a given self-reported behavior measured in Wave 3. The independent variable is whether the participant gave the correct answer in the memory task about the given behavior. Dependent variables are standardized: coefficients are expressed as SD.

Table A3.4: Does remembering the norm predict policy support? Policies supported by majority

	Regulation on red meat (1)	Fossil fuel rationing (2)	Rental tax on poor insulation (3)	Ban on cars in city center (4)
Correct recall	**0.212*****	**0.247*****	**0.295*****	**0.230****
Controls	Yes	Yes	Yes	Yes
Observations	303	303	303	303

* p<0.05, ** p<0.01, *** p<0.001.

Results from OLS regressions with weighted data. The dependent variable is support for a given policy measured in Wave 3. The independent variable is whether the participant gave the correct answer in the memory task about the given policy. Dependent variables are expressed as binary variables with values 0 and 1. Coefficients, hence, can be interpreted as probabilities.

Table A3.5: Does remembering the norm predict policy support? Policies supported by minority

	VAT on meat (1)	Toll on highways (2)
Correct recall	-0.037	**-0.334*****
Controls	Yes	Yes
Observations	303	303

* p<0.05, ** p<0.01, *** p<0.001.

Results from OLS regressions with weighted data. The dependent variable is support for a given policy measured in Wave 3. The independent variable is whether the participant gave the correct answer in the memory task about the given policy. Dependent variables are expressed as binary variables with values 0 and 1. Coefficients, hence, can be interpreted as probabilities.

A3.5 Wave 2 questionnaire instructions

This appendix presents the instructions of our treatments and our main outcomes of interest from Wave 2.

A3.5.1 Attention check question

This is a question for us to check that you are paying attention to the survey. Please let us know that you are reading all our questions carefully by selecting "1" on the response scale.

[Slider ranging from 1 to 10]

Only participants who selected "1" on both attention check questions were allowed to continue the survey.

A3.5.2 Norms treatment

During Wave 1, we asked questions about three domains of everyday consumption (meat, heating, and car use).

In particular, we asked you, for each of these 3 domains, what are your usual level of consumption and what levels of consumption you consider to be "ethically appropriate", given their impact on the environment.

Now that we have analyzed the information collected in the first wave of this survey, we can inform you about the **behaviors and attitudes** of the majority*.

*The statistics we provide here are based on wave 1, and are adjusted to be representative of the Luxembourgish population in terms of age and gender.

What do we mean by "behaviors and attitudes of the majority"?
Those adopted by more than 50% of the population.

[next page]

Home heating
For the majority, when it is less than 10 ℃ outside:

- the temperature of the home is **20℃ or less.**
- the home temperature considered ethically appropriate is **20 ℃ or less.**

In other words, only a minority heat their homes and consider it appropriate to heat one's home above 20 ℃.

[next page]

Did you know? Only **decreasing home temperature by 1°C** can **decrease energy consumption by 3% to 10%** (Fraunhofer Institute for Buildings Physics, UK Energy Saving Trust).

[next page]

Meat, fish or seafood consumption
The majority:

- eat **6 meals** containing meat, fish or seafood per week **or less**.
- feel that it would be ethically appropriate to eat **4 meals** per week **or less.**

In other words, only a minority eat more than 6 meals containing animal proteins per week and consider appropriate to eat more than 4 meals per week.

[next page]

Car versus public transport for a regular trip

To switch from a 30-minute car trip to public transport (bus, train, tramway) or to soft mobility (walking, cycling, electric bikes, scooters), the majority:

- would be willing to increase their travel time **by 15 minutes or more**.
- would consider ethically appropriate to increase their travel time **by 20 minutes or more**.

In other words, only a minority would not be willing to put up with an extra 15 minutes and would consider the appropriate extra time to be lower than 20 minutes.

[next page]

Use of public transport and soft mobility

69% use either public transport (bus, train, tram) or (at least 10 minutes of) soft mobility **several times a week or daily**.

A3.5.3 Policy treatment

In wave 1, we asked you questions about 3 areas of everyday life (meat, housing and mobility). For each of these 3 areas, we presented 2 hypothetical policies (a tax and a regulatory restriction). Thus, in total, we presented to you 6 **hypothetical policies** and asked you if you would support their introduction by 2025. We explained that the tax revenues from these policies would be used by the government to develop action plans providing alternative solutions to the population.

On average, respondents **believe that none of the 6 policies would be supported** by the majority.

In fact, **4 out of the 6 policies** are **supported** by the majority*.

*The statistics we provide here are based on wave 1, but are adjusted to be representative of the Luxembourgish population in terms of age and gender.

Here are the 4 policies, and the **proportion** of people who supported them:

- A tax of 10% on the rental income of landlords who rent out accommodations with poor energy efficiency **(64%)**
- A strict regulation on red meat production **(63%)**
- A rationing on fossil energy sources **(53%)**
- A car ban in densely populated areas **(50%)**

[next page]

Here are the 2 policies that were not supported by the majority, and the proportion of people who supported them:

- 17% VAT on meat (37%)
- A 5-euro toll on Luxembourg's motorways (21%)

A3.5.4 Personal norms

If participants were not allocated to the Norms treatment:

In the first wave, we asked about 3 areas of current consumption (meat, heating and cars).

In particular, we asked you, for each of these 3 areas, what levels of consumption you considered to be "**ethically appropriate**", given their impact on the environment.

We would like to ask you this question again. There is no right or wrong answer, please express your opinion at the time without trying to remember what you answered in wave 1.

If participants were allocated to the Norms treatment:

In this second wave, we have revealed to you what the **majority** considers **ethically appropriate** in three domains of everyday consumption (meat, heating, and car use).

Now, we would like to ask you again what **you** consider as **ethically appropriate**.

There are no right or wrong answers, please express your opinion of the moment without trying to remember what you had replied in wave 1.

From a sustainability viewpoint, what do you think is the maximal temperature that it is **ethically appropriate** to set in a dwelling when it is less than 10 ℃ outside?

[Slider ranging from 15 to 30]

As in wave 1, we define "eating meat, fish or seafood" by having a meal that contains at least 50 grams of either of these animal proteins.

Are you uncomfortable with the grouping of meat, fish or seafood? (click)

Some respondents argued that the lack of distinction between these three types of animal proteins is too constraining and that there are significant differences between them. While this is true, please consider that this aggregation is based on the fact that all these types of animal proteins are, in various ways, detrimental to the environment when they are produced on an industrial scale.

From a sustainability viewpoint, how many meals per week containing meat, fish or seafood do you think it would be **ethically appropriate** to eat?

[Slider ranging from 0 to 21]

In the first wave, we presented you with a **hypothetical scenario** in which you regularly make a journey that takes **30 minutes by car.**

We asked you to consider the additional time you would be willing to devote to using **public transportation** (bus, train, tramway) and/or **soft mobility** alternatives (walking, cycling, electric bikes, scooters).

Based on this scenario, please complete the following sentence:

"In view of adopting a more sustainable lifestyle, it would be **ethically appropriate** for me to systematically switch to public transport or soft mobility as long as the **additional time** (in minutes, on top of the initial 30 minutes) does not exceed ... "

[text box to write]

A3.5.5 Policy support

If participants were not allocated to the Policy treatment:

In wave 1, we asked you questions about 3 areas of everyday life (meat, housing and mobility). For each of these 3 areas, we presented 2 hypothetical policies (a tax and a regulatory restriction). Thus, in total, we presented to you 6 **hypothetical policies** and asked you if you would support

their introduction by 2025. We explained that the tax revenues from these policies would be used by the government to develop action plans providing alternative solutions to the population.

We would like to ask you again if you would support these policies under these circumstances.

If participants were allocated to the Policy treatment:

In wave 1, we presented to you 6 **hypothetical policies** and asked you if you would support their introduction by 2025. In this survey we revealed the results to you.

We would like to ask you again if you would support these policies.

For **housing**, the 2 hypothetical policies were:

- A **rationing on fossil energy** sources (for instance, a quota on heating oil or gas, depending on the size of the household)
- A **tax** of 10% on the rental income perceived by **landlords** who rent out **accommodations with insufficient energy efficiency** (EPC rating of D or worse).

These two policies would facilitate the implementation of a **massive housing plan to improve energy efficiency** (click)

- short-term support to reduce energy consumption
- free heating system servicing (tune-up and maintenance inspection)
- personalized support (at home or online) to change habits and reduce consumption
- support to structural renovation and equipment investments
- free energy audits to determine EPC rating and identify necessary investments
- need-based subsidies on insulation investments and renewable energy equipment
- zero interest rates on renovation and energy-efficient equipment

Under this massive housing plan...

Would you support the **rationing on fossil energy** sources?

[Yes/No]

Would you support a **tax** of 10% on the rental income perceived by **landlords** who rent out **accommodations with insufficient energy efficiency**?

[Yes/No]

For **meat**, the 2 hypothetical policies were:

- stricter **regulations** on **red meat** production (e.g. only allowing meat from cull cows that have reached the end of their milk production or breeding function, stricter environmental protocols in farms,. . .)
- increasing the **Value Added Tax rate (VAT)** on meat, fish or seafood to **17%** (current rate is 3%)

These 2 policies would facilitate the implementation of a **a massive plan to promote vegetarian alternatives** (click)

- Vegetarian alternatives are made more visible than meat in supermarkets stalls
- Vegetarian alternatives are subsidized (price reduced by half)
- A label comparing the environmental impact per protein intake of meat and of vegetarian alternatives
- A label comparing the price per protein intake of meat and of vegetarian alternatives

- Free consultations with nutritionists to ensure that vegetarian meals contain all necessary nutrients

Under this massive meat substitution plan...
Would you support the **reduction** of the available quantity of **red meat** in supermarket stalls, butcheries and restaurants?
[Yes/No]
Would you support an increase of the **Value Added Tax rate (VAT)** on meat, fish or seafood to **17%**?
[Yes/No]

For **mobility**, the 2 hypothetical policies were:

- a **car ban** in densely populated areas
- a **5-euro toll** for every use of Luxembourgish highways.

These two policies would facilitate the implementation of a **massive mobility plan based on public transport and soft mobility** (click)

- Giving higher priority to buses in bottlenecks
- Developing the soft mobility infrastructure (e.g., more bike lanes, more rental bikes,...)
- Reduce delays / uncertainty about trip duration
- Reduce the duration of trips (fewer connections, provide more public transport hubs)
- Make hubs and public transport safer
- Increase provision in order to prevent overcrowding (increase frequency of services, expand hours of services early morning and late evening)

Under this massive mobility plan...
Would you support the **car ban** in densely populated areas?
[Yes/No]
Would you support the **5-euro toll** on Luxembourg's motorways?
[Yes/No]

A3.5.6 Intended behaviors

In wave 1 you stated that the usual temperature at your home is [response from Wave 1] ℃ when it is less than 10 ℃ outside.
In the upcoming month, what is the likelihood that you will change your house temperature when it is less than 10 ℃?

- 0 - Completely certain of not changing
- 1 - Unlikely to change
- 2 - 50-50
- 3 - Likely to change
- 4 - Completely certain of changing

If participants did not select 0 in the previous question:

What do you think will be the **usual temperature** of your dwelling when you are at home when it is less than 10℃, **in the upcoming month**? (in ℃)

[Slider ranging from 15 to 30]

In wave 1 you stated that, over the last 7 days, you ate [response from Wave 1] meals containing meat, fish or seafood (out of 21).

In the upcoming 3 months, what is the likelihood that you will change your meat, fish and seafood consumption?

- 0 - Completely certain of not changing
- 1 - Unlikely to change
- 2 - 50-50
- 3 - Likely to change
- 4 - Completely certain of changing

If participants did not select 0 in the previous question:

How many meals containing meat, fish or seafood do you think you will eat per week in the upcoming 3 months?

[Slider ranging from 0 to 21]

In wave 1, you stated that you use public transport (bus, train, tram) at a frequency of "[response from Wave 1]" and soft mobility (walk, cycle or use a (electric) scooter for a trip of 10 minutes or more) at a frequency of "[response from Wave 1]".

In the upcoming 3 months, what is the likelihood that you will change your use of **public transport**?

- 0 - Completely certain of not changing
- 1 - Unlikely to change
- 2 - 50-50
- 3 - Likely to change
- 4 - Completely certain of changing

If participants did not select 0 in the previous question:

How frequently do you think you will use public transport (bus, train, tram) in the upcoming 3 months?

- Daily
- Several times a week
- Several times a month
- Several times a year
- Never

In the upcoming 3 months, what is the likelihood that you will change your use of **soft mobility**?

- 0 - Completely certain of not changing
- 1 - Unlikely to change
- 2 - 50-50
- 3 - Likely to change
- 4 - Completely certain of changing

If participants did not select 0 in the previous question:

How frequently do you think you will use soft mobility (walk, cycle or use a (electric) scooter for a trip of 10 minutes or more) in the upcoming 3 months?

- Daily
- Several times a week
- Several times a month
- Several times a year
- Never

A3.6 Wave 3 questionnaire instructions

This appendix presents the instructions of our main outcomes of interest from Wave 3.

A3.6.1 Attention check question

This is a question for us to check that you are paying attention to the survey. Please let us know that you are reading all our questions carefully by selecting "1" on the response scale.

[Slider ranging from 1 to 10]

Only participants who selected "1" on both attention check questions were allowed to continue the survey.

A3.6.2 Self-reported behaviors

In the same spirit as the first two waves, we would like to ask you briefly about your **recent practices** related to home life, meat consumption, and transportation **over the past two months**.

There are no right or wrong answers. For the benefit of the study, please try to answer with complete honesty.

Over the last 7 days, **how many meals** containing meat, fish or seafood did you eat?

[Slider ranging from 0 to 21]

How frequently have you been using public transport (bus, train, tram) over the past two months?

- Daily
- Several times a week
- Several times a month
- Once or twice over the past two months
- Never

How frequently have you been **walking, cycling** or using a (electric) **scooter** for a trip of 10 minutes or more, over the past two months?

- Daily
- Several times a week
- Several times a month
- Once or twice over the past two months
- Never

A3.6.3 Policy support

In the previous waves, we asked you whether **you would support 6 hypothetical policies** in the domains of housing, food consumption and mobility. For each domain, tax revenues from these policies would be used **to fund massive public plans** to improve citizens' access to sustainable

solutions. We would like to ask you one last time whether **you would support** the following policies.

- A rationing on fossil energy sources
- A 10% tax on rental income from energy-inefficient properties
- A strict regulation on red meat production
- 17% VAT on meat
- A 5-euro toll on Luxembourg's motorways
- A car ban in densely populated areas

[Yes/No]

A3.6.4 Treatment memory tasks

If participants were allocated to the Norms treatment in Wave 2:

In the second wave of the survey, we shared **information about the sustainability-related behaviors and attitudes** of other participants. **Do you remember?**

Below are **3 statements** based on that information. Please indicate whether each statement is **true** or **false**.

Please note, your responses will **not affect your compensation**, but we appreciate your **best effort** to recall.

- For the majority of respondents, their home temperature is **21°C or higher** when the outside temperature is below 10 °C.
- The majority of participants **eat** at least **8 meals** containing meat, fish or seafood per week.
- **Less than 50%** of participants use either public transport or soft mobility **several times a week or daily.**

[True/False]

If participants were allocated to the Policy treatment in Wave 2:

In wave 2 we revealed to you **which policies were supported by a majority** in wave 1. **Do you remember** which?

Please note, your responses will **not affect your compensation**, but we appreciate your **best effort** to recall.

Please mark the policies that were supported by a majority of participants in wave 1.

- A rationing on fossil energy sources
- A 10% tax on rental income from energy-inefficient properties
- A strict regulation on red meat production
- 17% VAT on meat
- A 5-euro toll on Luxembourg's motorways
- A car ban in densely populated areas
- I really don't remember

Conclusion

Conclusion

This dissertation investigates the role of social incentives in shaping prosocial behavior. Specifically, I focus on two kinds of social incentives, social esteem and social norms, and their relationship with COVID-19 vaccination uptake and proenvironmental behaviors. Across three chapters, I offer both theoretical and empirical contributions to our understanding of how individuals respond to social incentives, how accurately they perceive these incentives, and how such incentives can be leveraged in public policy.

Chapter 1 addresses a key theoretical claim in the literature on image crowding-out effects: that monetary incentives may reduce prosocial behavior by undermining the social esteem associated with it. To test this empirically, we introduce a novel, incentive-compatible methodology to measure social esteem based on vignettes and incentivized second-order beliefs. Applied in the context of COVID-19 vaccinations, our experiment reveals that rewards can significantly reduce social esteem; however, we do not find this effect for penalties. These results are consistent with our theoretical framework. Beyond contributing to the literature, this work provides a practical tool that policymakers can use to predict when monetary incentives might crowd out prosocial behavior.

In Chapter 2, using incentivized second-order belief elicitation, we show that individuals substantially and systematically underestimate how sustainable others are by misperceiving their behaviors, personal norms, and policy support. These underestimations are consistent with both pluralistic ignorance and false consensus effects. Furthermore, we demonstrate that social expectations predict not only behavior and policy support but also personal norms, suggesting that social norms may serve as levers for behavioral change.

Building on this, Chapter 3 tests whether these misperceptions can be corrected and whether doing so leads to lasting changes in behavior, personal norms, and policy support. In a longitudinal experiment spanning nine months, we implemented norm-based information treatments to correct the misperceptions identified in Chapter 2. We show that our treatments can increase sustainable behaviors, personal norms, and support for restrictive policies. These effects are driven by positive belief updating and upward behavioral adjustment.

Taken together, the three chapters of this dissertation offer a unified contribution to the study of social incentives and prosocial behavior. By combining theory and empirical testing, I aim to contribute both to the academic literature and to policymaking by providing informed guidance to practitioners.

Future Research Directions

This dissertation opens several avenues for future research. Perhaps the one I find most interesting is extending my studies to contexts of heterogeneous norms and polarization (te Velde 2022). Society is increasingly transitioning toward a state where there may not be a single, universal normative view of what constitutes correct or morally acceptable behavior; rather, a pluralism of normative views has emerged (Panizza *et al.* 2024). Political polarization can also influence beliefs about what is socially desirable (Bursztyn, Egorov and Fiorin 2020).

Behaviors that some groups consider prosocial, such as maintaining a vegetarian diet, engaging in environmentally friendly actions, or receiving a COVID-19 vaccine, may not be viewed as prosocial or desirable by others. For instance, Bénabou and Tirole (2006)'s framework on incentives and prosocial behavior assumes a common normative view within society of what constitutes prosocial behavior. However, this assumption may not hold in many contexts. People who hold strong anti-vaccine attitudes, for example, may not perceive someone who takes the vaccine as more prosocial; rather, they may see the opposite (Gallegos, de Castro Pecanha and Caycho-Rodríguez 2023, Mylan and Hardman 2021). This can occur even if anti-vaxxers do not actively view vaccination as harmful to society per se, but because vaccination may signal to them that the individual does not belong to their group and therefore does not share their general normative beliefs (Iyengar *et al.* 2019, Moore-Berg *et al.* 2020).

If we were to extend the studies in this thesis to contexts where social incentives pull in opposite directions for different groups, surprising findings might emerge. The recent U.S. elections, for example, provide a unique opportunity to explore several polarized norms, such as climate change mitigation behaviors, vaccination uptake, marijuana use, and discrimination against vulnerable groups (Panizza *et al.* 2024). In such contexts, Democrats and Republicans may react very differently to the introduction of monetary incentives aimed at promoting or discouraging these behaviors. Likewise, the effectiveness of norm-based interventions may depend heavily on the specific reference group being targeted: right-wing supporters may be indifferent to the norms of left-wing groups, and vice versa.

In a society with increasingly polarized norms, policymakers and scholars need to understand the interplay between political identities, ideologies, narratives, and social incentives, and how these factors, in turn, influence behavior. Recognizing that populations are not only becoming more heterogeneous but also increasingly divided by opposing beliefs is not just essential but perhaps unavoidable for the future of social research and policymaking.

References

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and prosocial behavior." American economic review, 96(5): 1652–1678.

Bursztyn, Leonardo, Georgy Egorov, and Stefano Fiorin. 2020. "From extreme to mainstream: The erosion of social norms." American economic review, 110(11): 3522–3548.

Gallegos, Miguel, Viviane de Castro Pecanha, and Tomás Caycho-Rodríguez. 2023. "Anti-vax: the history of a scientific problem." Journal of Public Health, 45(1): e140–e141.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. "The origins and consequences of affective polarization in the United States." Annual review of political science, 22(1): 129–146.

Moore-Berg, Samantha L, Lee-Or Ankori-Karlinsky, Boaz Hameiri, and Emile Bruneau. 2020. "Exaggerated meta-perceptions predict intergroup hostility between American political partisans." Proceedings of the National Academy of Sciences, 117(26): 14864–14872.

Mylan, Sophie, and Charlotte Hardman. 2021. "COVID-19, cults, and the anti-vax movement." The Lancet, 397(10280): 1181.

Panizza, Folco, Eugen Dimant, Erik O Kimbrough, and Alexander Vostroknutov. 2024. "Measuring norm pluralism and perceived polarization in US politics." PNAS nexus, 3(10): pgae413.

te Velde, Vera L. 2022. "Heterogeneous norms: Social image and social pressure when people disagree." Journal of Economic Behavior & Organization, 194: 319–340.

List of Figures

List of Tables