UNIVERSITY OF CALIFORNIA

Santa Barbara

Motivational systems in dyadic cooperation are designed for reputation-based partner choice

A dissertation submitted in partial satisfaction of the

requirements for the degree

Doctor of Philosophy

in

Psychological & Brain Sciences

by

Sakura Arai

Committee in charge:

Professor Leda Cosmides, Chair

Professor James Roney

Professor Daniel Conroy-Beam

Professor John Tooby

June 2022

The dissertation of Sakura Arai is approved.

_____

Professor James Roney

_____

Professor Daniel Conroy-Beam

_____

Professor John Tooby

_____

Professor Leda Cosmides, Committee Chair

June 2022

Motivational systems in dyadic cooperation are designed for reputation-based partner choice

ACKNOWLEDGEMENTS

I would like to express my sincerest gratitude to my advisors, Leda Cosmides and John Tooby, for their enduring support and guidance throughout my years at UCSB. Without their wisdom, patience, and kindness, I would not have made it through the worst of times during my studies. In particular, I am forever indebted to Leda for being willing and able to go above and beyond her "pay grade" to help me out when I needed it most.

I am grateful to my committee members, Jim Roney and Dan Conroy-Beam, for their insights and constructive suggestions. I would like to thank the Fulbright Program as well as the Yoshida Scholarship Foundation for their financial and organizational support. Thanks should also go to my colleagues at the CEP and in the department who have shown me friendship. I would also like to acknowledge the undergraduate research assistants for their hard work in collecting the data presented here. Lastly, I owe a lot to my mother for always being there for me.

I would like to dedicate this dissertation to the memory of my late advisor, Toshio Yamagishi. I would not be here today if it were not for him. I will be ever grateful to him for having trust in me and for leaving a monumental legacy that continues to provide inspiration.

Curriculum Vitae

Sakura Arai

June 2022

Education

Ph.D., Psychological & Brain Sciences — June 2022 (expected)
University of California, Santa Barbara,

M.A., Psychology — 2018
University of California, Santa Barbara

M.S., Life Sciences — 2015
University of Tokyo

B.A., Cognitive and Behavioral Sciences — 2013
University of Tokyo

Publications

Arai, S., Tooby, J., & Cosmides, L. (2022). Motivations to reciprocate cooperation and punish defection are calibrated by estimates of how easily others can switch partners. *PLOS ONE*, 17(4): e0267153.

Spadaro, G., Graf, C., Jin, S., **Arai, S.**, Inoue, Y., Lieberman E., Rinderu, M. I., Yuan, M., van Lissa, C., & Balliet, D. (2022). Cross-Cultural Variation in Cooperation: A Meta-Analysis. *Journal of Personality and Social Psychology*.

Eisenbruch, A. B., Lukaszewski, A. W., Simmons, Z. L., **Arai, S.**, & Roney, J. R. (2018). Why the wide face? Androgen receptor gene polymorphism does not predict men's facial width-to-height ratio. *Adaptive Human Behavior and Physiology*, *4*(2), 138-151.

Inoue, Y., Takahashi, T., Burriss, R. P., **Arai, S.**, Hasegawa, T., Yamagishi, T., & Kiyonari, T. (2017). Testosterone promotes either dominance or submissiveness in the Ultimatum Game depending on players' social rank. *Scientific reports*, *7*(1), 5335.

Wu, J., Balliet, D., Tybur, J. M., **Arai, S.**, Van Lange, P. A. M., & Yamagishi, T. (2017). Life history strategy and human cooperation in economic games. *Evolution and Human Behavior*, 38(4), 496–505.

Awards

Arizona State University Interdisciplinary Cooperation Winter School Travel Grant — 2020

Fulbright Scholarship — 2015 – 2020

Yoshida Scholarship, Overseas Study Program — 2016 – 2019

Graduate Student Association Travel Grant, UCSB — 2017, 2020

International Travel Award, Japanese Society of Social Psychology — 2015

Dean Prize, College of Arts and Sciences, University of Tokyo — 2013

Distinguished Young Researcher Poster Award — 2012
The 5th Annual meeting of the Human Behavior and Evolution Society of Japan

ABSTRACT

Motivational systems in dyadic cooperation are designed for reputation-based partner choice

by

Sakura Arai

Dyadic cooperation is the building block of human social exchange. But forming cooperative partnerships poses two problems: choosing partners and being chosen by partners. A growing body of research suggests that reputation-based partner choice creates competition to be chosen by desirable cooperative partners and therefore motivates people to behave generously and acquire a reputation as a valuable cooperator.

However, a reputation as a cooperator may also attract cheaters, undesirable partners who do not reciprocate cooperation. Evidence indicates that inflicting punishment can deter cheating, but appearing punitive may drive away cooperators as well. Despite the apparent dilemma, little is known about motivations to punish in the presence of competition to be chosen. I hypothesize that motivational systems are designed to attract desirable partners and manage reputations by up-regulating cooperation and down-regulating punishment behaviors; systems will do so in response to cues of reputation-based partner choice—cues indicating that one is in competition to be chosen as a partner.

Three studies tested this hypothesis. Studies 1 and 2 assessed motivations to cooperate and punish using economic games with a punishment option. Cues of reputation-based partner choice were either measured (as estimates of how many outside options potential partners would have; study 1) or experimentally manipulated (in study 2; cues of group membership and anonymity indicated whether one is being evaluated as a potential partner).

Study 3 examined whether there is a trade-off between acquiring cooperative versus punitive reputations.

Results support the hypothesis. The cues of reputation-based partner choice up-regulated motivations to cooperate while down-regulating motivations to punish. It is also shown that punishing harms one's cooperative reputation and lowers the probability of attracting partners, confirming that the function of these motivational calibrations is to improve one's reputation as a cooperation partner. The present research provides evidence that motivational systems are designed for managing reputations to be chosen by desirable cooperation partners.

TABLE OF CONTENTS

Chapter 1: Introduction 1
1.1 Reputations as mental representations 2
1.2 Reputation-based partner choice 3
1.3 Managing reputations for partner choice 4
1.3.1 Managing reputations for attracting desirable partners 5
1.3.2 Managing reputations for deterring undesirable partners 7
1.3.3 Costs of acquiring a reputation as a punisher 9
1.4 The current dissertation 11
Chapter 2: Study 1. Motivational regulations based on opportunities for partner choice 13
2.1 Evolutionary models of dyadic cooperation 13
2.1.1 Strategies for cooperation under different conditions: social ecologies with low versus high partner choice 15
2.1.2 The problem of being chosen 17
2.1.3 Estimating degrees of partner choice 20
2.1.4 Relational mobility reflects partner choice in a social ecology ..23
2.1.5 The current experiment 25
2.2 Study 1 Methods 27
2.2.1 Participants 27
2.2.2 Design 28
2.3 Study 1 Results 32
2.3.1 What predicts the decision to switch partners? 32
2.3.2 Inflicting punishment 41
2.3.3 What predicts reciprocation? 46
2.3.4 Are qualities of the person used to update priors based on social ecology? 52
2.4. Study 1 Discussion 53
2.4.1 Evidence that motivational systems are designed for social ecologies with varying levels of partner choice 53
2.4.2 What is the function of punishment in dyadic reciprocal cooperation? 57
2.4.3 Micro and macro effects of social ecology 59
2.4.4 Limitations and future directions 62
2.4.5 Conclusions 63
Chapter 3: Study 2. Motivational regulations based on reputation concern 65
3.1 Study 2 Introduction 65
3.1.1 Motivations to cooperate when you are being evaluated as a partner 65
3.1.2 Motivations to punish when you are being evaluated as a partner? 72
3.1.3 The current experiments 74
3.2 Study 2a: Prolific sample, online 74
3.2.1 Study 2a Methods 74
3.2.2 Study 2a Results 80
3.2.3 Study 2a Discussion 84
3.3 Study 2b: College sample, online 86
3.3.1 Study 2b Methods 87
3.3.2 Study 2b Results 88
3.3.3 Study 2b Discussion 92

3.4 Study 2c: College sample, in-person 94
3.4.1 Study 2c Methods 94
3.4.2 Study 2c Results 96
3.4.3 Study 2c Discussion 99
3.5 Study 2 Pooled-analysis 101
3.5.1 Pooled-analysis Methods 101
3.5.2 Pooled-analysis Results 101
3.5.3 Pooled-analysis Discussion 103
3.6 Study 2 General discussion 104
3.6.1 Ingroup-favoring cooperation as a byproduct of reputation management 104
3.6.2 Is out-group discriminating punishment a product of reputation management? 106
3.6.3 The targets of cooperative and punitive reputations 109
3.6.4 Conclusion 110
Chapter 4: Study 3. The trade-off between cooperative and punitive reputations ... 112
4.1 Study 3 Introduction 113
4.2 Study 3a 118
4.2.1 Study 3a Methods 118
4.2.2 Study 3a Results 120
4.2.3 Study 3a Discussion 125
4.3 Study 3b 126
4.3.1 Study 3b Introduction 126
4.3.2 Study 3b Methods 128
4.3.3 Study 3b Results 129
4.4 Study 3 General discussion and conclusions 135
Chapter 5: General discussion 139
5.1 Evidence that motivational systems are designed for reputation-based partner choice 139
5.2 Reputation-based partner choice may select for punitive motivations 141
5.3 Limitations and future directions 144
5.3.1 A general skepticism on manipulating reputation concern by situational cues 144
5.3.2 Withdrawing cooperation: motivations to do so and its reputational consequences 145
5.4 Conclusion 147
References 148
Appendix A 173
Appendix B 194
Appendix C 212

Chapter 1: Introduction

From the moment of birth, humans rely on cooperation from others. Human infants are altricial—they cannot live without care from adults—and it takes about two decades before children can calorically support themselves (Hill & Hurtado, 2009). But in ancestral settings, becoming able to sustain oneself does not mean there is no more reliance on others. People count on each other for sharing food and other necessary resources in everyday life (Bliege Bird et al., 2012; Gurven, Allen-Arave, et al., 2000; Kaplan et al., 1985). Most notably, without having someone else who is willing to help, one cannot survive recurrent reversals in life such as illness, injury, and bad luck in hunting and gathering (Sugiyama, 2004). These selection pressures are likely to have shaped our adaptations to seek and cultivate cooperative partnerships (Tooby & Cosmides, 1996).

Dyadic social exchange, where two parties reciprocally deliver benefits to each other, is the most basic and ubiquitous form of human cooperation (Tooby et al., 2006). Because forming and retaining such partnerships is based on a mutual agreement, it involves two problems: choosing partners and being chosen by partners. First, one needs to recognize potential partners and evaluate who would be the most beneficial partners (e.g., are they interested in forming a relationship with me? Are they willing to deliver benefits to me? Are they able to?). Second, because others also prefer to form partnerships with the most beneficial partners available to them, one needs to demonstrate how valuable one would be as a partner so that others would agree to form a partnership. This dissertation examines how our mind solves the second problem of being chosen as a cooperation partner, specifically,

how human motivational systems are designed to regulate behaviors for attracting desirable partners.

1.1 Reputations as mental representations

Based on what kind of information do we choose a partner? Let us call the kind of information used to assess a partner "*reputation*". A reputation is a belief about a specific aspect of an individual (Barclay, 2015; Yamagishi & Matsuda, 2003). A basic form of reputation would be about a trait—e.g., a reputation for being cooperative, wise, hostile, formidable, tall, good at swimming—or a combination or summary of several traits—e.g., reputation as an easy prey, a reputation as a tenacious foe. By definition, a reputation exists only as a mental representation held by an individual. Yet, as the everyday use of the word *reputation* implies, a reputation can be shared among multiple individuals (e.g., everyone thinks Alex is a fast runner) and can be represented as a shared representation (e.g., "I think everyone thinks that Alex is a fast runner") (Barclay, 2015; Sperber & Baumard, 2012).

A reputation does not necessarily reflect reality because it is formed based on limited information an individual gathers through personal experience, observation, or information from other individuals. An individual who you think is generous might be seen as stingy by your friend, and both reputations can be accurate (e.g., the individual likes you and is generous only to you) or inaccurate (e.g., neither you nor your friend have enough experience with the individual). In other words, one can attain a reputation that does not reflect one's true characteristics by changing behaviors in the presence of (different) observers.

An individual can be ascribed multiple *reputation**s***, not just *a* reputation (e.g., Alex has a reputation for being generous, and separately, Alex also has a reputation as an early-riser and a reputation for having bad eyesight). Naturally, some reputations of an individual are closely related to how desirable the individual is seen as a cooperation partner, some are not (but can be relevant to different kinds of evaluation, e.g., desirability as a mate, undesirability as an enemy).

1.2 Reputation-based partner choice

Before reviewing what kind of reputations are important for the problem of being chosen as a cooperation partner, let me first sketch out where these reputational evaluations emerge. When individuals have a freedom to choose a partner, they exert *partner choice*: They choose the most desirable partner from a pool of partners available to them. Then there emerges a market-like competition for access to the most desirable partners, which is known as a *biological market* (Noë & Hammerstein, 1994, 1995). In biological markets, individuals choose partners based on values potential partners would offer (Noë & Hammerstein, 1994, 1995). This means that partner choice necessitates reputations—or however you call it, beliefs about individuals as potential partners (otherwise, there is no real choice—partners would be forming pairs randomly or indiscriminately). The term *reputation-based partner choice* emphasizes that partner choice is exercised based on reputations, suggesting that the

best way to attract beneficial partners is to invest in reputations (Roberts, 1998; Roberts et al., 2021). [1]

1.3 Managing reputations for partner choice

Organisms can invest in—or manage—their reputations by influencing how other individuals mentally represent them, thereby changing the probability of acquiring desirable partners. Although it pays an organism to be able to infer how others represent its own reputations (Leimar & Hammerstein, 2001), investing in reputations does not necessarily require Theory of Mind or mind-reading capacities (Barclay, 2015; Manrique et al., 2021). It does not even require an ability to mentally represent one's own reputations—recognizing the presence of an audience and changing behaviors would suffice.

In fact, organisms that appear not to possess mind-reading capacities manage their reputations by changing their behaviors in the presence of potential partners. For example, cleaner wrasses, *Labroides dimidiatus*, form a mutualistic relationship with their "client" fish by eating ectoparasites and dead skin from clients, but cleaners can "cheat" by feeding on client tissue, a more desirable food source for cleaners (Grutter & Bshary, 2003). However, when other potential clients are around, cleaner fish behave more cooperatively (i.e., eat less tissue) than when there are no observers (Bshary & Grutter, 2006; Pinto et al.,

[1] There are differences between mathematical formalizations of reputation-based partner choice and biological markets. For example, models of the former take place as two games, a game for reputation-building and a game for partner choice, while models of the latter do not (Roberts et al., 2021). These differences are not relevant to how reputation-based partner choice is conceptualized in this dissertation. Also see Roberts et al. (2021) for how reputation-based partner choice is modeled differently from indirect reciprocity.

2011). Moreover, cleaners serve low-value clients to encourage them to stay; the apparent function of this behavior is to attract desirable clients, which prefer to interact with cleaners that other clients have chosen (Bshary, 2002). Cleaners living in a high-competition environment with many potential clients and rival cleaners employ this strategy more often than those in a low-competition environment with few clients and rivals (Binning et al., 2017). These findings illustrate that reputation management can be achieved by a mere tendency to adjust motivations, e.g., to cooperate, when there are potential partners.

1.3.1 Managing reputations for attracting desirable partners

What kind of reputations are relevant to being chosen as a cooperation partner? In humans, there is ample evidence that individuals who display willingness and ability to deliver benefits are preferred as reliable cooperative partners (Barclay, 2013, 2016). Lab experiments and field data demonstrate that people prefer to associate with individuals who demonstrate their willingness to provide benefits to others (Barclay & Willer, 2007; Bliege Bird & Power, 2015; Feinberg et al., 2014; Gurven, Allen-Arave, et al., 2000; Sylwester & Roberts, 2010, 2013). Studies also suggest that individuals with the ability to confer benefits are sought as cooperation partners (Eisenbruch et al., 2016; Eisenbruch & Roney, 2017), although people regard ability to provide as less important than willingness to do so (Bliege Bird & Power, 2015; Eisenbruch & Roney, 2017). Displaying willingness to cooperate and provide generously seems to be the most straightforward strategy for investing in one's reputation as a cooperation partner.

As the example of cleaner fish shows, a strategy for managing reputation can be as simple as a tendency to adjust one's motivation to cooperate when potential partners are

present. Humans too employ reputation management strategies like this. People become more generous when they believe that there are others observing them, regardless of whether observers actually exist (Bradley et al., 2018). Even children as young as three years old manage their cooperative reputation by behaving generously when their behaviors are observed (Buhrmester et al., 1992; Kelsey et al., 2018; Leimgruber et al., 2012; Z. Wu et al., 2018). This tendency to display one's generosity and fairness is shown to increase over childhood (Shaw et al., 2014). In humans, it appears that up-regulating motivations to behave generously is the primary strategy for investing in one's reputation as a valuable cooperator.

However, to attract the most desirable partners, up-regulating motivations to cooperate in the presence of an audience might not be enough. Because valuable partners are often sought out by others, there are "outbidding" competitions to appear *more* cooperative than others (Noë & Hammerstein, 1994; Roberts, 1998). The competition to be chosen becomes more intense as the number of competitors—alternative options for desirable partners—increases (Baumard et al., 2013; Debove et al., 2015). The presence of competitors indeed up-regulates motivations to display willingness to provide: People behave increasingly generously as there are more competitors they need to outbid (Barclay & Willer, 2007; Debove et al., 2015). These findings suggest that motivational systems take various inputs indicating the presence of competition to be chosen and in response up-regulate motivations to cooperate.

1.3.2 Managing reputations for deterring undesirable partners

Acquiring a reputation as a cooperator may be a mixed blessing. It can attract not only desirable partners who reliably reciprocate but also undesirable ones who intend to take advantage of your cooperativeness. Cheating—a failure to reciprocate cooperation—is an inevitable obstacle in a pursuit of a reciprocal relationship (Trivers, 1971). Undesirable partners may be hard to avoid because they sometimes "fake" generosity to be chosen and then start cheating (or under-reciprocating) once partnerships are formed (Barclay & Willer, 2007; Bshary, 2002).

One solution may be investing in another reputation—one that would discourage those who intend to cheat—thereby increasing the probability of forming relationships only with reliable cooperators. Acquiring a reputation as a punisher is an example. Punishment—imposing a cost that reduces the payoff of a cheater—is shown to suppress cheating (Fehr & Gächter, 2000, 2002; Yamagishi, 1986), and those who are recognized as punishers deter selfish behaviors (dos Santos et al., 2013). Similarly, individuals known for being willing and able to inflict costs can deter offenses in general, and people, young men in competition especially, pursue these reputations (Cohen & Nisbett, 1996; Daly & Wilson, 1988). Moreover, there is empirical evidence that motivations to punish are designed to deter mistreatment by cheaters and bystanders (Delton & Krasnow, 2017; Krasnow et al., 2016; Yamagishi et al., 2009).

Acquiring a reputation for being willing and able to punish may not only deter cheating but also increase the benefits one receives from social exchanges. To avoid the risk of getting punished, not only cheaters but also cooperators might up-regulate motivations to cooperate and over-reciprocate when interacting with those who are known to punish. Plus,

having the ability to inflict costs—being formidable—alone can give you leverage over your partners and increase the weight they place on your welfare relative to theirs (Sell et al., 2009). Thus, investing in a reputation for being willing and able to inflict costs would be more advantageous when your partners appear not to value your welfare relative to theirs (Lim, 2012), such as when they do not reciprocate.

Additionally, motivations to punish might be simply selected for by lowering the fitness of other individuals. Punishment, inflicting a cost, can increase one's payoff relative to others (Price et al., 2002). As a strategy to achieve a competitive advantage, punishment is expected to be more efficient against cheaters than cooperators because it will be perceived as more legitimate and therefore will be less likely to invite retaliation (Raihani & Bshary, 2019). Systems may be designed to up-regulate motivations to punish those who did morally wrong such as cheating, where inflicting a cost can be legitimized as a means of deterring further wrongdoings.

However, very little research has been conducted to examine how motivations to punish are regulated when there is competition to be chosen. dos Santos et al. (2013) show that people up-regulate motivations to punish a stingy partner when other potential partners can observe their punishment behaviors—i.e., when they can acquire punitive reputations. But, because partner choice was not allowed in this study, it is unclear whether people would do so when they can leave a stingy partner and switch to a generous one. Relatedly, several studies show that people punish cheaters more when they believe that others are observing their behavior (Batistoni et al., 2022; Kamei, 2018; Kurzban et al., 2007; Piazza, 2008). These findings indicate that the presence of potential partners—which might include undesirable ones who attempt to cheat—up-regulate motivations to punish. They offer

tentative evidence that motivational systems are designed to invest in a punitive reputation to deter undesirable partners.

1.3.3 Costs of acquiring a reputation as a punisher

Nevertheless, appearing too punitive may defeat the purpose of managing reputations to be chosen by desirable partners. First, not all "cheating" is intentional or dispositional. Even those who intend to cooperate sometimes fail to reciprocate by accident (Delton et al., 2012). Strictly punishing these unintended failures can make a partner unwilling to cooperate with you, resulting in a vicious cycle of mutual defection (Delton et al., 2011). When partner choice is possible, the partner may even leave you for a more forgiving partner who does not punish.

Moreover, being known as a punisher can drive away potential cooperative partners, even if they are not the direct target of punishment. Considering that anyone can fail to reciprocate by mistake, it is risky even for cooperators to interact with those who tend to punish any apparent "cheating", which inevitably includes innocent mistakes. Plus, an error can be made by punishers as well in judging behaviors. Even if you know that you are a reliable cooperator and will not fail to reciprocate, those who tend to punish may misjudge your behaviors or intentions and punish you by chance.

Second, inflicting punishment can indicate potentially unfavorable traits as a dyadic cooperation partner such as aggressiveness, competitiveness, and dominance. Regardless of whether punishment is directed toward cheaters, the act of punishing—inflicting a cost—is an aggression (Clutton-Brock & Parker, 1995). Even if it is clear to observers that punishment is legitimate (e.g., it is in response to an intentional cheating), punishers could

be seen as having antagonistic and uncooperative tendencies. These traits would make punishers undesirable as partners in peaceful dyadic exchanges.

Thirdly, there are nuances in negative sanctions. There are more subtle and amicable ways to reduce the payoff of a cheater than punishing, such as withdrawing or withholding the benefits of cooperation from the cheater (e.g., TIT FOR TAT; Axelrod & Hamilton, 1981), leaving the cheater for a more cooperative partner (Hammerstein & Noë, 2016), and verbal communication to or about the cheater (e.g., reproach, gossip) (Guala, 2012; Molho et al., 2020). Using language, people can even negotiate with under-reciprocating partners—without necessarily inflicting a cost—and tell them to up-regulate the levels of cooperation. When these options are available, those who choose to punish are likely to be viewed as less forgiving and more antagonistic than those who employ other methods.

Indeed, studies show that those who inflict punishment are generally less preferred over those who do not as partners in reciprocal partnerships (Dhaliwal et al., 2021; Horita, 2010; Ozono & Watabe, 2012). Acquiring a reputation as a punisher can even hurt your reputation as a cooperator. In public goods games, those who punish cheaters are sometimes rated as less cooperative than non-punishers (Kiyonari & Barclay, 2008; Mifune et al., 2020).[2] These findings imply that punishment is a double-edged sword for reputation management, and that regulating punitive motivations is a delicate balancing act of deterring undesirable partners by acquiring a reputation as a punisher while attracting desirable partners by acquiring a reputation as a cooperator.

[2] But see Barclay (2006). Also, third-party punishers—those who punish cheaters who have cheated *others* in dyadic cooperation—are evaluated more favorably than non-punishers (Dhaliwal et al., 2021; Jordan et al., 2016; Nelissen, 2008; Raihani & Bshary, 2015b). Detailed discussions are found in Chapter 4.

1.4 The current dissertation

Existing evidence suggests that the mind may be sensitive to situational cues suggesting reputation-based partner choice is relevant, and that these cues serve as inputs to motivational systems regulating behaviors in dyadic social exchange. I hypothesize that motivational systems are designed to attract desirable partners and manage one's reputations by calibrating cooperation and punishment behaviors. This predicts that cues of reputation-based partner choice should up-regulate motivations to cooperate and invest in one's reputation as a reliable cooperator. But do the same cues affect motivations to inflict punishment? Despite an apparent dilemma between regulating motivations to invest in one's cooperative versus punitive reputations, it is unexplored how these motivations are regulated together. If there is a trade-off between acquiring cooperative versus punitive reputations, the same cues of reputation-based partner choice should down-regulate motivations to punish to protect one's reputation as a cooperator.

These predictions were tested in studies 1 and 2 of this dissertation. Economic games with a punishment option were used to measure motivations to cooperate and punish. Studies 1 and 2 each tested effects of two cues of reputation-based partner choice. In study 1, one cue was an "internal" estimate of how many outside options potential partners have, and the other was an "external" situational cue of whether switching to another partner was allowed in the experiment. Study 2 manipulated two situational cues suggesting that one is being recognized and evaluated as a potential partner.

I conducted an additional vignette study to investigate if the reputational trade-off exists. Study 3 examined whether punishing a cheater harms one's reputation as a cooperator and lowers the probability of chosen by others.

Chapter 2: Study 1. Motivational regulations based on opportunities for partner choice [3]

2.1 Evolutionary models of dyadic cooperation

The evolution of dyadic cooperation has been explored through evolutionary game theory since the 1970s. A consistent finding is that decision rules that cause cooperation can evolve and be maintained in a population by natural selection if agents can implement a strategy for *conditional cooperation*. These are strategies that direct benefits to agents who cooperate rather than those who defect. Defectors—*cheaters*—are individuals who accept the benefits of cooperation but fail to provide sufficient benefits in return, either by not reciprocating at all or by reciprocating too little (Trivers, 1971). There are, however, many different strategies for conditional cooperation. Which ones are favored by selection depends on the social ecology—especially on the extent to which it provides options for switching partners.

In early models of conditional cooperation—also known as *reciprocity*—agents were not permitted to choose partners or to avoid defectors by switching partners. Agents were randomly paired with their partners, and they interacted with each partner repeatedly. They could recognize and remember (at least some of) their history of interaction with a given

[3] This work has been published as: Arai, S., Tooby, J., & Cosmides, L. (2022). Motivations to reciprocate cooperation and punish defection are calibrated by estimates of how easily others can switch partners. PLos One. 17(4): e0267153.

partner, and use that information to decide whether to cooperate or defect in a given round. This social ecology favored *sanction-based strategies*, such as TIT FOR TAT, which cooperates when their partner delivers benefits and defects when that partner defects (Axelrod & Hamilton, 1981). These strategies are stable against invasion by strategies that defect because they respond to defection by withholding benefits or inflicting costs ("punishment"), and they resume cooperation only after the defecting partner cooperates again. When agents with sanction-based strategies are paired with other conditional cooperators, they repeatedly harvest the benefits of mutual cooperation, allowing them to outcompete strategies that defect. Because switching partners to avoid defectors is not an option in these models, Hammerstein and Noë (2016) call them "partner control models without outside options." In this social ecology, cooperation is maintained by natural selection because agents monitor their partner's behavior and "control" it through positive and negative sanctions. Empirical work suggests that some non-human organisms use sanction-based strategies in reciprocal cooperation (Bshary & Grutter, 2002; Dugatkin & Alfieri, 1991; Schweinfurth & Taborsky, 2020).

Sanction-based strategies differ in detail: For example, some leave defectors (Aktipis, 2004; Hayashi, 1993; Izquierdo et al., 2010; Joyce et al., 2006; Li et al., 2021; Schuessler, 1989), some cooperate contingently, withdrawing cooperation after one defection (Axelrod & Hamilton, 1981), and yet others require several defections, thereby maintaining cooperation with conditional cooperators who defected by mistake (Axelrod, 1984). But one's reputation as a cooperator, defector, or punisher plays no role in these strategies, beyond the history of interaction remembered by one's current partner.

In the 1990s, evolutionary scientists began to explore selection in *biological markets*: social ecologies in which agents can leave one cooperative partner and choose another (Bull & Rice, 1991; Noë, 1990; Yamagishi et al., 1994). These *partner choice models* assume that agents can infer and represent the reputation of multiple potential partners based on available information, such as their behavior when interacting with other individuals (did they cooperate? defect? punish?) or other observable traits (e.g., skill procuring valued resources). They also assume that agents can use reputation information in deciding whether to stay with their current partner or switch to a different one. In these models, competition to be chosen—or retained—as a cooperative partner "controls" defection and stabilizes cooperation by the threat of partner switching (Hammerstein & Noë, 2016). Partners who defect are abandoned for partners who are more likely to provide benefits.

A social ecology in which agents can switch partners favors *reputation-based strategies*: ones that (i) prefer partners who are likely to reciprocate—ones with a reputation as a reliable cooperator, and (ii) manage their reputation to attract valuable cooperative partners (Leimar & Hammerstein, 2001; Nowak & Sigmund, 1998; Ohtsuki & Iwasa, 2006). Empirical studies have shown that many organisms, including humans, behave as if they have evolved reputation-based strategies (Barclay & Willer, 2007; Bshary & Grutter, 2002; Simms et al., 2006).

2.1.1 Strategies for cooperation under different conditions: social ecologies with low versus high partner choice

The strategies favored by selection differ in these two contexts because they pose quite different adaptive problems, especially regarding the best response to defection (Barclay &

Raihani, 2016; Baumard et al., 2013; Martin & Cushman, 2015). The early models, which prevent partner choice entirely, are the most extreme version of a *low partner choice ecology*. In this social ecology, the only way to minimize the costs of defections is to sanction the defecting partner. If the partner is not reciprocating at all, one can go on strike—refuse to provide benefits until the partner starts to cooperate—or punish the defection by inflicting a cost on the partner, possibly at some cost to oneself. Neither party realizes the benefits of mutual cooperation until the partner responds by cooperating. If the partner is under-reciprocating, one can down-regulate the benefits one provides successively, until the partner responds by providing more in return. But one does not have the option of switching to a more rewarding partner.

Selection pressures are different in a *high partner choice ecology*. The most extreme version is a social ecology in which many alternative cooperative partners are available, information about their reputations is free, and there is no cost to switching partners. Under these conditions, the opportunity cost of staying with a partner who defects or under-reciprocates is high. An *opportunity cost* is the benefit one would gain by choosing the best alternative option; in this case, the opportunity cost is equal to the benefits you would harvest by interacting with the most cooperative alternative partner *who is willing to interact with you*. The opportunity cost is high when the payoff of remaining with a partner who defects or under-reciprocates is lower than the payoff of switching to a more cooperative partner.

High opportunity costs select against sanction-based strategies—even those that never pay a cost to punish a defector. When you down-regulate or withdraw cooperation to reform an uncooperative partner, you are forgoing the benefits of mutual cooperation that you could

gain by interacting with a different, more cooperative partner. In a high partner choice ecology, abandoning your current partner for a more cooperative one is more fitness-promoting than retaining and trying to reform an uncooperative partner. This is true even if your current partner does reciprocate; selection favors switching partners when your best outside option provides higher payoffs than your current partner.

2.1.2 The problem of being chosen

Switching to a new, more cooperative partner will not be an option, however, if high value cooperative partners do not want to interact with you. Because valuable cooperative partners will prefer to interact with the most rewarding partners available to them, developing a reputation for cooperation is a way of competing for good partners in ecologies where partner choice is high (Barclay, 2013; Roberts, 1998). But what kind of reputation will attract valuable cooperative partners?

2.1.2.1 Reputation for providing benefits

The most straightforward way to acquire a reputation as a good cooperator is to resist temptations to cheat and behave cooperatively (Baumard et al., 2013). Enhancing this reputation can be accomplished by providing as much—or more—than others in your social ecology (Barclay & Willer, 2007); initiating cooperative relationships by delivering benefits (Quillien, 2020); or demonstrating skill at acquiring resources (Eisenbruch et al., 2016). People invest in acquiring a cooperative reputation, even in the laboratory: They are more generous in cooperative games when they can be observed by third parties (Bradley et al., 2018). And partner choice can elicit "competitive altruism": When the observer will have

the opportunity to choose a cooperative partner, people are more generous than when partners are fixed or randomly assigned (Barclay, 2004; Barclay & Willer, 2007; Sylwester & Roberts, 2010, 2013).

2.1.2.2 Reputation for inflicting negative sanctions?

A reputation for sanctioning failures to reciprocate may deter defection whether partner choice is low or high. But does it harm your reputation as a valuable cooperator in high partner choice ecologies?

Not all failures to reciprocate arise from a disposition or intent to profit from the temptation to cheat. An otherwise good cooperator can make a mistake or be temporarily unable to reciprocate due to injury or lack of resources (Delton et al., 2012). Under these circumstances, sanctioning a failure to reciprocate can trigger defection in return, jeopardizing the flow of benefits that result from mutual cooperation (Delton et al., 2011). Sanctioning mistakes carries additional risks when partner choice is high: An otherwise good cooperator may leave you for a partner who is less punitive and more rewarding. Sometimes there is a downside to sanctioning even intentional defections: An occasional defector who provides higher net benefits than any of your outside options may leave for a more forgiving partner.

When partner choice is high, imposing negative sanctions not only risks a current relationship; it could threaten future ones as well. There can be reputational costs to withdrawing benefits and, especially, to inflicting punishment (Raihani & Bshary, 2015a). Few studies directly compare the effects of these two methods of sanctioning in cooperative interactions. But the reputational consequences of punishing have been explored in a handful

of studies in which participants witness several potential partners who vary in how punitive they are toward others. When asked if they wanted to interact with a specific partner in various economic games, participants were less likely to choose punitive over non-punitive partners as recipients (Dhaliwal et al., 2021; Horita, 2010; Ozono & Watabe, 2012), although punishers were sometimes more likely to be preferred as providers (Horita, 2010) (but see Ozono & Watabe, 2012). Potential partners who sanctioned by punishing had a worse reputation than those who sanctioned by rewarding (Dhaliwal et al., 2021; Ozono & Watabe, 2012). In another study, punishers were trusted less (and proved less trustworthy) than non-punishers, whereas generous behavior elicited trust (Przepiorka & Liebe, 2016).

Taken together, these studies suggest that inflicting punishment can decrease one's desirability as a potential cooperative partner. In high partner choice ecologies, this reputational cost may not be compensated by eliciting more cooperation from defectors than withdrawing benefits does, at least when strangers interact. In a repeated prisoners dilemma (PD) in which there were two methods for sanctioning a partner—inflicting punishment or withdrawing for one round—punishment did not elicit more cooperation than withdrawing cooperation (Barclay & Raihani, 2016).

In sum, what counts as adaptive behavior varies with social ecology. When partner choice is limited, the only way to elicit cooperation from an uncooperative partner is to withhold benefits or inflict punishment. But these negative sanctions may be unnecessary—and possibility counter-productive—in high partner choice ecologies, where one's bargaining power depends on having good outside options: alternative partners who are not only cooperative, but also willing to choose you. These considerations raise an under-

explored question: Does information about partner choice in one's local ecology calibrate motivations to cooperate and punish?

2.1.3 Estimating degrees of partner choice

Computational systems that generate motivations to reciprocate, defect, or punish regulate cooperative behavior. Their evolved design should reflect selection pressures common in the social ecologies of our group-living hominin ancestors. Did these social ecologies select for designs that implement sanction-based strategies or reputation-based strategies?

The evolutionary models discussed above represent two extremes on a partner choice continuum. At one extreme are models in which one can either engage in a relationship with a single partner or forgo cooperation entirely. At the other extreme are models in which many cooperation partners are available and switching partners is cost-free. But neither extreme was common during hominin evolution.

Hunter-gatherers are rarely forced to engage with one and only one cooperative partner, even when they live in very small bands. They usually have the option to forage individually rather than cooperatively, or to cooperate exclusively with kin (which does not require reciprocation to be advantageous) (Bliege Bird et al., 2012; Gurven, Hill, et al., 2000; Sugiyama, 2004). Nor did they have access to an unlimited number of partners with zero cost of switching. Most social ecologies were intermediate between these two extremes.

Does this imply that human motivations to reciprocate, defect, or punish are tuned to a social ecology with a single, intermediate level of partner choice? Not necessarily. The availability of cooperative partners, i.e., the pool of potential partners for dyadic

cooperation, depended, in part, on band size, which varied with foraging conditions from ~25 men, women, and children—2 to 3 extended families—to as many as 500 for more sedentary hunter-gatherers and for nomadic bands when they periodically aggregate (Kelly, 2003). This variation could occur within a lifetime (with changes in season, rainfall, and game dispersal) and over generations: From the first appearance of anatomically modern humans, global climate has alternated between ice ages and warming periods; sometimes changed by 10° C (18° F) within a few decades (Ziegler et al., 2013); and varied with latitude as hominins dispersed across the globe. We propose that this variation selected for motivational systems that treat partner choice as a continuous variable and adjust behavior accordingly. All else equal, the perception that other people can easily switch partners should up-regulate motivations to reciprocate their help and down-regulate motivations to sanction their defections.

This calibration requires mechanisms that can estimate the degree of partner choice in the situation one is facing. This can be decomposed into two questions: (i) How much partner choice is there in my local social ecology *in general*, and (ii) what are the prospects for partner switching *right now*, in my immediate situation? Estimating the probability that one's current partner can switch to a better outside option is a judgment made under uncertainty. A mechanism that is well-designed for estimating this probability might implement a Bayesian updating process (Delton et al., 2011; Pietraszewski et al., 2015).

When you have no previous history with a new person, and no specific knowledge about that person's value as a cooperator, the prior probability that this person will be able to switch to a better outside option should be based on estimates of partner choice in your local social ecology. This estimate reflects the prospects for switching partners for a person

randomly drawn from that ecology. It can be based on a variety of cues, such as how many individuals you encounter on a regular basis, how frequently you encounter new people, how easy it is to change social groups, how trustworthy the average person is, the prevalence of exploitive behavior (including violence), whether the environment is resource-rich or resource-poor, and the number of individuals who can afford to share resources with others.

This prior can be updated based on cues present in the immediate situation. These cues might speak to qualities of the *person* or features of the *situation*. Qualities of the person relevant to their outside options are judged from thin information: In ultimatum games, participants who see photos of their partners' faces offer more to those whose faces had been rated (by others) as more attractive, kind, cooperative, healthy, trustworthy, higher in status, and (surprisingly) more productive as a cooperative forager (Eisenbruch et al., 2016, 2019). A prior based on social ecology could be updated based on features of the situation as well: Are we temporarily isolated or are alternative partners available right now (Debove et al., 2015)? Does this situation draw people from my ingroup or an outgroup (Yamagishi et al., 1999)? A mechanism that is well-designed for estimating the probability that a specific partner will switch should use person-specific and situation-specific cues to adjust a prior based on social ecology upward or downward. The resulting estimate—a posterior probability—represents your current partner's ability to leave you for a better partner, compared to an average individual from your local ecology.

This posterior probability should reflect both the local social ecology and cues about the immediate situation. When cues in the immediate situation are minimal, the posterior probability should be closer to the prior probability, which was based on the local social ecology. As you gain more experience of a particular partner, the posterior probability may

depart more from that prior. In either case, the posterior probability that your current partner can easily switch partners should calibrate your motivations to cooperate with that person and invest in your reputation as a valuable cooperator.

2.1.4 Relational mobility reflects partner choice in a social ecology

For a given social ecology, what is the prior probability that a newly encountered individual will have the opportunity to leave a cooperative partnership with you to form a new one? This probability is proportional to *relational mobility*: the number of opportunities in a given society for individuals to form new relationships (Yuki et al., 2007). The more such opportunities the average person has, the greater the degree of partner choice in that society.

A twelve-item scale created by Yuki, Schug, and colleagues (Yuki et al., 2007) measures people's perceptions of relational mobility. It first prompts the rater to think of others in their immediate society, such as people in their workplace or neighborhood. It then asks how much the rater agrees with statements about other people, such as "They have many chances to get to know other people," "There are few opportunities for these people to form new friendships" (reverse-scored), "If they did not like their current groups, they would leave for better ones." The relational mobility scale measures the extent to which other people are seen as having many alternative partners to choose from in the context specified (e.g., group members, friends, or other relationships).

Perceptions of relational mobility vary across societies: Average scores are higher in the US than in Japan, for example (Thomson et al., 2018). These scores predict societal differences in motivations that are theoretically relevant to partner choice. In their review of

the literature on relational mobility, Yuki, Schug, and colleagues summarized how people in different societies react to incentives created by levels of relational mobility (Oishi et al., 2015; Yuki & Schug, 2020). In societies where people believe relational mobility is high, they are geared toward (i) looking for new partners and evaluating their qualities, as well as (ii) advertising one's qualities as a partner and displaying commitment to desirable partners (Komiya et al., 2019; Schug et al., 2010; Thomson et al., 2018). These behavioral tendencies suggest the operation of reputation-based strategies: efforts to choose better partners based on their reputation and to be chosen by improving one's reputation as a valuable cooperator.

Where relational mobility is low, people behave as if they have few outside options. Oishi et al. (Oishi et al., 2015) report that people in these social ecologies are more likely to (i) invest in maintaining cooperation within small, close-knit groups, and (ii) avoid being excluded from these close-knit cooperative relationships by cooperating rather than defecting with their current partners. They behave as if their partners are enacting sanction-based strategies: They cooperate with existing partners by default, and assume their partners are ready to respond to defection by imposing negative sanctions (Yamagishi et al., 1999, 2008).

The prior probability that other people in your social ecology can switch partners varies with relational mobility: the number of opportunities the average person has to form new relationships. *Perceptions* of relational mobility are mental representations of this prior probability: They reflect the mind's estimate of how much partner choice others can exercise in your local social ecology. If the mind is designed to treat partner choice as a continuous variable, then measures of relational mobility should regulate motivations to reciprocate, defect, and punish, especially when interacting with people you do not know.

2.1.5 The current experiment

Study 1 investigates the design of motivational systems that regulate dyadic cooperation. The goal is to see if motivations to cooperate with, punish, and/or switch partners are calibrated by estimates of the degree to which others can exercise partner choice. If the mind is designed to treat partner choice as a continuous variable, these motivations should vary with relational mobility—an estimate of partner choice in one's local social ecology—and with verbal cues, delivered with the instructions, about partner choice in the immediate situation. As estimates of the probability that a partner can switch increase, we expect concern with one's reputation as a cooperative partner to increase, leading to more reciprocation and less punishment.

The experiment proceeded as follows. To measure motivations to cooperate with, punish, and switch partners, we used a game from behavioral economics in which two individuals can benefit by mutual cooperation: a Trust Game with Punishment (TGP) (Krasnow et al., 2012). It provides two interacting individuals—a truster and a responder—an opportunity to benefit each other by reciprocally cooperating (see figure 1.1). The truster, who starts with 100 points, decides how many to invest in their partnership. Because the invested points are tripled, both partners can be better off, but only if the responder shares enough of them with the truster (more than 1/3). If the responder gives too little, the truster has an opportunity to punish this decision (see section 2.2.2.1).

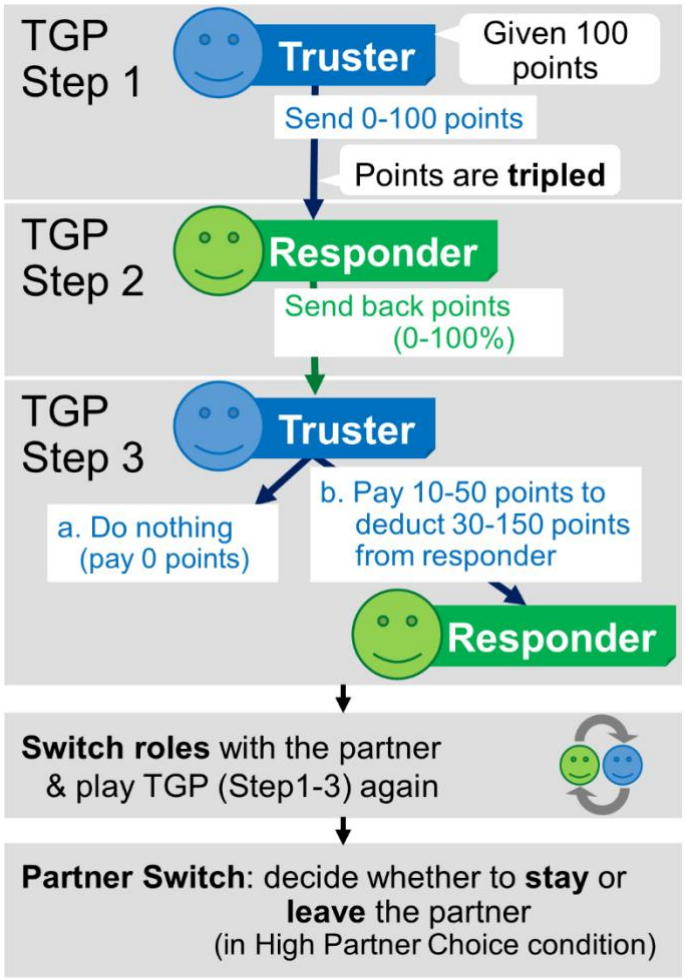


Figure 1.1. The flow of TGP and partner switching. Participants interacted with their partner in the Trust Game with Punishment (TGP). After interacting once (as the truster or the responder), the participant and the partner switched roles and interacted in the TGP again (order counterbalanced). After interacting with the same partner twice, once in each role, participants in the High Partner Choice condition decided whether they wanted to continue interacting with their current partner in the next TGP, or switch to a new partner. Participants in the Low Partner Choice condition were reminded that they would continue interacting with the same partner.

Each participant completed the relational mobility scale, to measure their estimate of how much partner choice others in general can exercise in their local social ecology. Before the TGP, half were told they could switch partners after 2 rounds of the game (*High Partner Choice* condition), and half were told they would interact with the same partner for the

entire study (*Low Partner Choice* condition). These instructions served as cues to partner choice in the immediate situation. After two rounds, participants in the former condition were asked whether they wanted to switch partners.

In addition to measuring individual perceptions of relational mobility, the study was conducted in two countries where perceptions of relational mobility differ on average: the US (high) and Japan (low).

2.2 Study 1 Methods

2.2.1 Participants

Participants (N = 1039) were from the US (n = 519, 53.5% male, M_{age} = 39, SD_{age} = 12), recruited via Amazon Mechanical Turk, and Japan (n = 520, 53.8% male, M_{age} = 42, SD_{age} = 10), recruited via an equivalent crowd-sourcing website, Lancers, with instructions translated to Japanese by a native speaker (the first author). They were compensated approximately 3 dollars (either in US dollars or Japanese Yen) for their participation in the study, which lasted about 25 minutes.

Those who wished to participate in the study first completed an informed consent form. After the study, participants received a written debriefing about the study design and purposes. They were then asked for consent to use their data; it was explained that they would be compensated regardless of their answer. Fourteen participants who did not provide consent were excluded from the analysis. This study was approved by the Institutional Review Board at University of California, Santa Barbara (Human Subjects Committee). See Appendix A for materials.

2.2.2 Design

There were two experimental conditions: High versus Low Partner Choice. Participants were randomly assigned to one of them. After reading instructions for a TGP (Krasnow et al., 2012), they were told that they would be paired with a partner.

In the Low Partner Choice condition, participants were told that they would be interacting with the same partner for the rest of the study. In the High Partner Choice condition, the instructions explained that, after interacting with the same partner twice in two TGPs, they had a choice: (i) They could switch to a new, unknown partner, or (ii) they could remain with their current partner for the next TGP—*but only if that partner chose to remain with them*. Thus, they knew before their first TGP that keeping their current partner might depend on their reputation in that partner's eyes.

Note that participants in the High Partner Choice condition could decide they wanted to interact with a new partner after the second round, but they were not permitted to choose among alternative partners (and had no information about such partners). This was for strict experimental control, to ensure that the High and Low conditions differed in only one respect: whether people could leave their current partner or not.

2.2.2.1 Reciprocation and punishment in the TGP

Before the TGP, participants were told that they were going to be given points that could be used during the interaction. They were asked to imagine that the points they earned would be converted to real money at the end of the study. In the TGP, participants experienced both roles, truster and responder (order counterbalanced across participants). The TGP was same as the standard Trust Game, except a punishment phase was added after

the responder's decision (Krasnow et al., 2012). We will use terms such as "reciprocation" and "punishment" to describe the logic of the game, but these terms were not used in the instructions to participants.

Participants were told they would interact with another participant. In reality, they interacted with sham partners simulated by a program. This procedure, which was the only deception in the study, was necessary to examine hypotheses about how people react to different reciprocal behaviors.

Before the interaction began, each participant (real and sham) was given 50 points as "a bonus." This was done to ensure that trusters had enough points to punish the responder, regardless of how many points the responder returned to the truster.

The TGP had the following structure. The truster was given an endowment of 100 points to send or keep. The truster could send any number of points to the responder, from $0 \leq P \leq 100$, in 10-point increments. The P points sent to the responder were tripled, and the responder decided what percentage of (now) 3P points to return to the truster. (Options were displayed as both percentages of 3P and points; see Appendix A.) The percentage returned is the dependent variable that measures reciprocation by the participant (Dependent Variable [DV] 1: *Reciprocation by the participant*; 0-100% in 10% increments).

If the truster sends nothing to the responder, the truster keeps all 100 points. Sending points is a risky *investment*—because they are tripled, both parties can be better off, but only if the responder sends enough points back to the truster. Any points sent to the responder are at risk because the responder could decide to send nothing back to the truster, or so few points that the truster is worse off than if she had *not* risked the P points that she invested. The truster, whose payoff is $[(100 – P) + (.X \times 3P)]$, breaks even when the responder returns

1/3 of 3P points (payoff = 100, i.e., 100 – P + P). The truster realizes a positive payoff when the responder returns more than 1/3 of the tripled points and incurs a loss when less than 1/3 of 3P is returned.

After seeing what percentage the responder gave to the truster, the truster had the option to pay 10 points to subtract 30 points from the responder; the truster could pay up to 50 points, in 10-point increments, to subtract up to 150 points from the responder. Note that the instructions referred only to subtracting points; this was not labelled "punishment". The instructions included examples to make sure that participants understood the consequences of various decisions (see Appendix A for the full text of instructions).

When the participant was the responder, the (sham) truster always sent 70 points to the participant (70% of the endowment). These were tripled to 210 points. The participant responded by deciding what percent of these points to return to the truster. If the participant returned less than 50% of the 210 points, there was a 50% chance that the truster would pay 20 points to deduct 60 points from the participant. Participants who returned 50% or more of the points they received were never punished. (Punishing cooperators—*anti-social punishment*—is a rare response in real life for these populations (Herrmann et al., 2008; Shinada et al., 2004). Our interest herein is motivations to cooperate with or leave partners who punish acts that could be perceived as failures to reciprocate sufficiently.)

When the participant was the truster, the responder returned either 50% or 20% of the 3P points that the participant had made available. The participant then decided whether to deduct points from the responder. The number of points the participant paid to deduct points from the responder is the dependent variable that measures the participant's willingness to punish the partner (DV2: *Amount paid to punish the responder*; 0-50 points).

After the instructions for the TGP, participants had two practice rounds, once as the truster and once as the responder. They then answered five comprehension check questions about the TGP and their experimental condition (see Appendix A). About 2% of the initial participants (20 people) failed this check; these individuals did not progress to the TGP phase of the study.

2.2.2.2 Partner switching after the TGP

After interacting with their partners in the TGP twice—once as a truster and once as a responder—participants were reminded that they were going to play the TGP again. Participants in the Low Partner Choice condition were reminded that they would continue interacting with the same partner. Participants in the High Partner Choice condition were asked whether they would like to stay with their current partner or switch to a different partner (DV3: the decision to switch partners). Before deciding, they were reminded that they would keep the same partner only if both they and their partner chose not to switch. (*N.B.*: participants did not have to pay a cost to switch or to stay.) At the point when a third TGP was about to commence, all participants were told that the program had decided that there would be no further rounds of the TGP.

2.2.2.3 Measures

After the TGP, participants completed the relational mobility (RM) scale twice, in different forms: the original and a modified version (order counterbalanced). The original RM scale asked participants how many opportunities they think people around them have to find new partners (*RM others*): e.g., “It is easy for them to meet new people.” The modified

one asked the same questions, but about themselves (*RM self*). The *RM self* scale had the same 12 items as the original scale, except that words referring to others were replaced with words referring to oneself: e.g., “It is easy for me to meet new people." *RM self* was added to control for individual differences in perceptions of one’s own opportunities to find new partners, which need not correspond to estimates of the relational mobility of other people in one’s social ecology. We also recorded which society participants were from (US or Japan).

2.3 Study 1 Results

Data were analyzed using R 4.0.3 (R Core Team, 2020). We examined the effects of the experimental manipulation (High versus Low Partner Choice condition), participants’ relational mobility scores (others and self), and society (US versus Japan) on the three DVs: (i) *Reciprocation by the participant* (DV1), (ii) *Amount paid to punish the responder* (DV2), and (iii) the decision to switch partners (DV3; only in the condition that permitted switching: High Partner Choice).

2.3.1 What predicts the decision to switch partners?

The logic of reputation-based strategies assumes that behavior in reciprocal interactions influences the probability that one’s partner will continue the cooperative relationship or switch to a different partner. So, we first examine whether decisions to reciprocate and punish affected DV3: the participant’s decision to switch partners. The opportunity to switch partners was available only in the High Partner Choice condition (n = 505).

Decisions to switch were made after the participant interacted with the same partner twice and experienced both roles: one interaction as truster, the other as responder. When given the option to stay or switch, 37.8% of participants chose to switch partners.

To determine which behaviors influence the decision to switch, we conducted logistic regressions, using the glm function in R (R Core Team, 2020). In preliminary analyses, we found that the order of roles—whether the participant played truster or responder first—did not predict decisions to switch partners, nor did the participants' relational mobility scores (others and self); *p*s > .05. These variables did not improve the Akaike Information Criterion (AIC) either, so they were removed from the model. The model focused on the choices participants made, the responses they experienced, and their society (US = 1, Japan = 0). We checked for multicollinearity using the Variance Inflation Factor (VIF values < 1.6). Although including interactions improved model fit, many of them showed strong multicollinearity even after centering variables by subtracting the mean (Robinson & Schumacker, 2009) (VIF > 10), making them difficult to interpret. For this reason, the model below does not include interaction terms.

Five predictor variables from the TGP were entered into the analysis. Two arise from the TGP in which the participant was the responder (and the [sham] truster sent P = 70 points):

- *Reciprocation by the participant*: What percent of 3P did the participant return to the truster? (0-100%).
- *Punishment received*: Did the truster punish the participant's response? (1 = punished, 0 = not punished).

Three predictors arise from the TGP in which the participant was the truster:

- *Trust*: How many points did the participant send to their partner? (P = 0-100 points).

- *Defection by the responder*: Did the (sham) partner respond by reciprocating (returning 50% of 3P) or defecting (returning 20% of 3P)? (50% = 0, 20% = 1).
- *Amount paid to punish the responder*: How much did participants pay to punish their partner's response? (0-50 points).

Figure 1.2 summarizes how each predictor affected the probability (adjusted odds ratio) that the participant would decide to switch partners when controlling for all the others. An odds ratio of 1 means the predictor variable had no independent effect on partner switching. See table 1.1 for the full model.

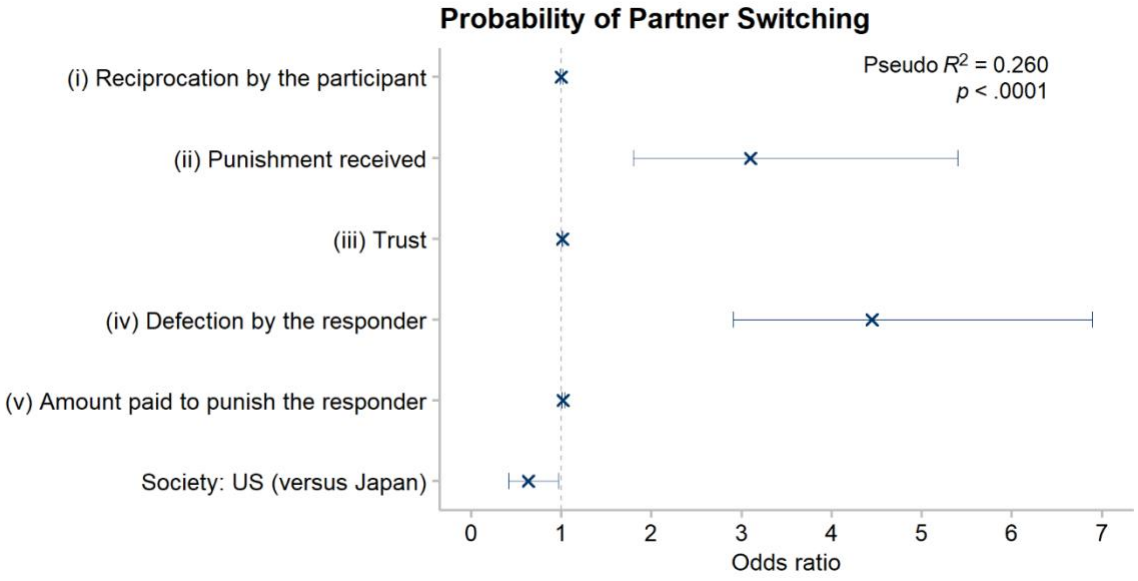


Figure 1.2. Adjusted odds ratio of each predictor for the decision to switch partners. Estimates of how much each predictor affected the decision to switch partners, when controlling for the five others. An odds ratio greater than 1 indicates a greater likelihood of partner switching; an odds ratio less than 1 indicates a lower probability of partner switching. Bars are 95% confidence intervals. *Reciprocation by the participant* = percent of 3P that the participant returned to the truster (0-100%). *Punishment by the partner* = whether the truster punished the participant's response (1, 0). *Trust* = P, the number of points the participant sent to the responder (0-100). *Defection by the responder* = the responder defected or reciprocated on the participant (1, 0). *Amount paid to punish the responder* = number of points the participant paid to punish the responder (0-50).

Table 1.1. Factors affecting the decision to switch partners.

Predictors	*b*	*SE*	Wald $\chi 2$	*OR*	95% CI	*p*
i. Reciprocation by the participant (0-100)	0.0009	0.01	0.02	1.00	[0.99, 1.01]	.883
ii. Punishment received (1, 0)	1.13	0.28	16.36	3.10	[1.80, 5.40]	< .001
iii. Trust (0-100)	0.01	0.00	11.02	1.01	[1.00, 1.02]	.001
iv. Defection by the responder (1, 0)	1.49	0.22	46.13	4.45	[2.91, 6.89]	< .001
v. Amount paid to punish the responder (0-50)	0.02	0.01	7.46	1.02	[1.01, 1.03]	.006
Society: US (vs. Japan)	-0.45	0.21	4.45	0.64	[0.42, 0.97]	.035

Note. Nagelkerke pseudo $R^2 = 0.26$. CI = confidence interval for *OR*. VIF values were < 1.6.

2.3.1.1 Are participants who were punished more likely to switch partners?

When participants were responders, the sham truster could punish them. Figure 1.2 shows the effect of each predictor variable on the decision to switch partners, when controlling for all the others. It shows that participants who were punished were three times more likely to switch partners than those who were not, controlling for the other predictors: Odds Ratio (OR) = 3.10 (95% CI = [1.80, 5.40]).

Because that odds ratio is based on all participants, it includes those who favored their partner over themselves by returning 50% or more (see table 1.2). Their decision to switch partners cannot be a response to being punished, however, because no one who returned 50% or more was ever punished. To see the effect of being punished on partner switching more clearly, the following analyses focus on participants who returned 40% or less, about half of whom were punished.

Table 1.2. Payoffs as a function of percent returned by the responder.

% returned	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Points returned to truster	0	21	42	63	84	105	126	147	168	189	210

Truster's payoff*	30	51	72	93	114	135	156	177	198	219	240
Responder's payoff	210	189	168	147	126	105	84	63	42	21	0

*The truster can earn 100 by not investing in the responder. Returning at least 40% gives the truster a positive payoff; 40% minimizes the difference in payoffs but favors the responder; 50% or more favors the truster over the responder.

Table 1.1 shows the payoffs to self (responder) and partner (sham truster) for each choice the responder could make. The sham truster always kept 30 points from the 100-point endowment; the 70 points the truster sent to the responder were tripled to 210 points, and the participant's task was to decide how many of these points to return to the truster. The responder's options were limited to 10% increments of 210. What counts as a failure to reciprocate, perhaps worthy of punishment?

Consider these payoffs in light of two concepts of reciprocity discussed in the literature: (i) ensuring the partner gains from having cooperated and (ii) ensuring equal payoffs for both partners (Baumard et al., 2013; Trivers, 1971). The sham truster could have earned 100 points (the endowment) by investing *nothing* in the responder; by risking 70 points, the sham truster enabled a positive payoff for both parties: the responder and self. A responder who returns more than 70 points—at least 40% of the tripled points—ensures a gain for the truster. Instead of 100 points, the truster will earn from 114 points (40% returned) to 240 points (100% returned).

Returning less than 70 points is a clear-cut case of defection. It creates a loss for the truster, who could have kept all 100 points, leaving the responder with nothing. Having risked 70 points, the truster takes a loss whenever the responder returns 30% (63 points) or less. The more the responder keeps, the worse off the sham truster is for having risked 70 of

100 points. This analysis is general to any positive number of points the truster sends. When the responder returns 40% (or more), the truster's payoff is positive: (100-P) + .4×3P = 100 + .2P (it would be positive for any return > [1/3]P). The truster's payoff is negative when the responder returns 30% or less: (100-P) + .3×3P = 100 - .1P.) Twenty-nine percent of responders returned 30% or less (148/505).

When participants were asked "How many points do you want to send back to your partner?", they chose from a display like the first two rows of table 1.1 (shaded in blue). It showed how many points the truster would receive when the participant returned X% of 210 points (see Appendix A). Because participants know that the truster risked 70 points, they know the truster realizes a net gain when more than 70 points are returned and a net loss otherwise. They also know the total payoff to the truster is the number of points the participant returns plus the 30 points that the truster kept.

No option results in equal payoffs for both of them. Equal payoffs would require returning 90 points to the sham truster—42.8% of 210— resulting in 120 for each (30 + 90 for truster, 210 - 90 for responder). Because responders were only allowed to return points in 10% increments, every option favors self over truster (40% or less) or truster over self (50% or more). A participant who views equality as appropriate reciprocation would return 40%: This option ensures a positive payoff for the partner while minimizing the difference in payoffs between self and truster (see below). The majority of responders (259/505 = 51%) chose options that bracketed 42.8% (strict equality) by returning 50% (180/505 = 36%) or 40% (79/505 = 16%). With these consequences in mind, we analyze responses to punishment.

Participants who returned 50% or more of the 210 points they received favored their partner, the sham truster, over themselves; they were never punished for this decision. Of these participants, 34% wanted to switch partners (95/278). The 227 participants who returned 40% or less favored themselves over their partner. Of those who were *not* punished, 31% wanted to switch partners (35/113)—comparable to the 34% found for those who returned 50% or more. So, in the absence of punishment, about one-third of participants decided to switch partners.

We next examine the effect of being punished on the 227 participants who returned 40% or less. These participants favored themselves over the sham truster, but to different degrees. Figure 1.3 shows how many participants returned from 0 to 100%, and the probability that they wanted to switch partners as a function of being punished by their current partner.

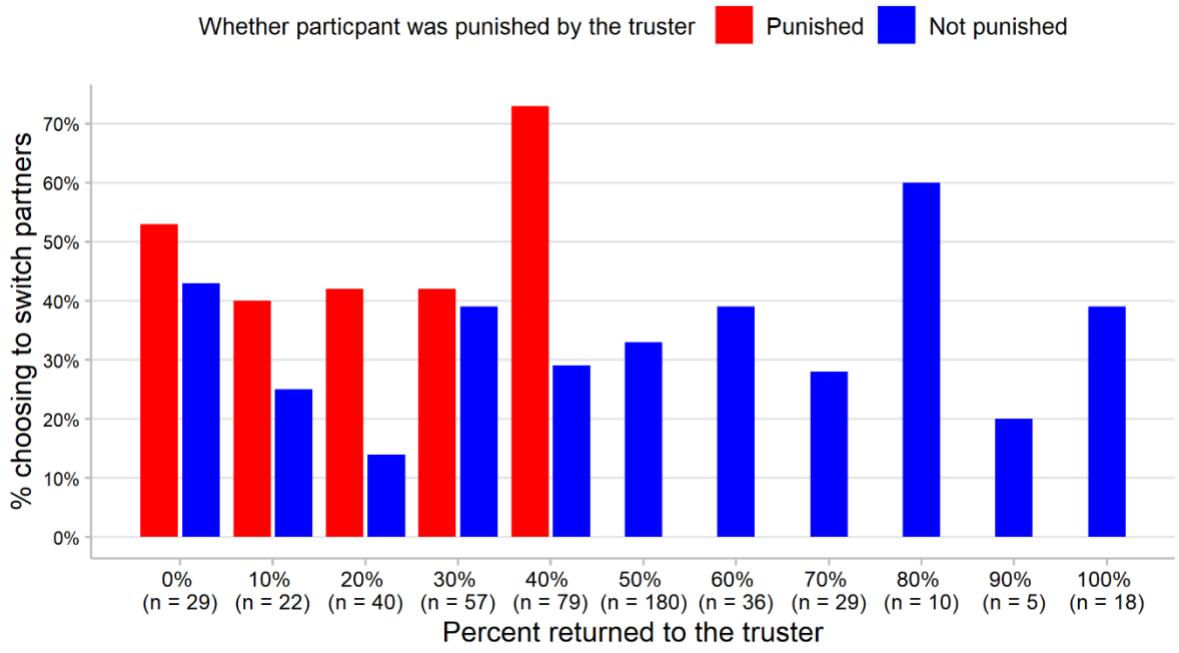


Figure 1.3. Probability of switching partners as a function of how much the participant returned and whether the participant was punished. The y-axis shows the percentage of participants who decided to switch partners in the High Partner Choice condition. The x-axis shows what percent of 210 points the participant (responder) returned to the partner (truster).

(How many individuals returned each amount is shown in parentheses.) Red bars: participants whose partner punished them by deducting 60 points; blue bars: participants who were not punished by their partner.

For the participants who returned 40% or less (n = 227), the probability of switching was higher for those punished than for those who were not punished: 53.5% versus 31% (61/114 vs. 35/113; Z = 3.44, p = .0003). They were 3.35 times more likely to switch when the truster punished them than not, controlling for all other predictors (OR = 3.35, 95% CI = [1.84, 6.27]).

What about unjust punishment? Those who returned 30% or less are clearly defectors, but is returning 40% a failure to reciprocate? These responders satisfied both concepts of reciprocity: 40% provides a positive payoff to both parties (114 points for truster, 126 points for responder) with the smallest deviation from equality—a 12-point difference. Returning 50%—105 points—is an equal division of the points the responder *received* but, because the truster kept 30 points of the initial endowment, it results in a larger departure from equality: a 30-point difference (135 points for truster, 105 points for responder). Do those who returned 40%—a positive payoff with almost equal outcomes—feel wronged by being punished?

Of the 79 participants who returned 40%, 73% of those who were punished wanted to switch partners (27/37), compared to 33% of those who were not punished (12/42). Controlling for other predictors, those who were punished for returning 40% were almost *ten times* more likely to switch partners than those who were not punished (OR = 9.58, 95% CI = [2.82, 40.81]). By contrast, being punished had no significant effect on the true

defectors—those who inflicted a negative payoff by returning 30% or less (n = 148) (OR = 1.99, 95% CI = [0.94, 4.33]).

2.3.1.2 Participants were more likely to stay with responders who reciprocated their trust.

As truster, the participant could send $0 \leq P \leq 100$ points to the responder. Sending $P > 0$ points creates 3P points, making cooperation for mutual benefit possible. The truster's payoff is 100 – P + .X(3P). Both benefit if the (sham) responder reciprocates by returning 50% of the 3P points: The truster gains a positive payoff because 100 – P + .5(3P) = 100 + .5P > 100. Returning 20% ensures a loss for the truster: 100 – P + .2(3P) = 100 - .4P < 100. This is a failure to reciprocate, indeed a defection (see above section 2.3.1.1). Participants were far more likely to leave defectors—partners who returned only 20% of 3P—than reciprocators (those who returned 50% of 3P): OR = 4.45 (95% CI = [2.91, 6.89]). Reciprocation by the responder greatly increased the probability that the participant wanted to continue their partnership.

2.3.1.3 Did participants who punished their partner want to remain in that relationship?

Switching partners defeats the purpose of punishing your *current* partner, if the function of punishing is to elicit more cooperation from a partner you plan to stay with (Axelrod & Hamilton, 1981). Yet those who punished their partner were not more likely to remain in that relationship; indeed, controlling for other predictors, the more points participants paid to punish the partner, the more likely they were to leave the punished partner (*Amount paid to punish the responder*: OR = 1.02; 95% CI = [1.01, 1.03]).

2.3.1.4 What else affected partner switching?

The more points participants entrusted to their partner, the more likely they were to want a new partner, although the effect was very small (*Trust*: OR = 1.01; 95% CI = [1.00, 1.02]). Also, American participants were less likely to switch partners than Japanese participants (*Society*: OR = 0.64; 95% CI = [0.42, 0.97]). How much participants returned to the truster was unrelated to their probability of switching partners (*Reciprocation by the participant*: OR = 1.00; 95% CI = [0.99, 1.01]).

2.3.1.5 Partner switching summary

Having been punished was the second largest predictor of the decision to leave one's partner in this study. Participants were less likely to stay with responders who defected—the largest predictor—and, all else equal, Americans were more likely to stay than Japanese participants. This implies that one's reputation as a cooperator affects the probability of keeping a partner: Reciprocation increases that probability and punishing decreases it. Participants' assumptions about relational mobility in their society did not predict their own decision to switch; their partner's behavior did.

2.3.2 Inflicting punishment

2.3.2.1 Did participants pay to punish?

When participants were the truster, they could punish their partner's response, whether the partner had returned 50% or 20% of 3P points. Twenty-seven percent of participants chose to inflict punishment (282/1039); 78% of these individuals were punishing defectors—those who returned 20% of 3P points (219/282). When the partner had defected,

44% of participants paid to punish the defection (219/496). Only 12% punished partners who had returned 50% of 3P points (63/543). (Trusters who punished reciprocators risked about 15 fewer points as truster than those who did not [P = ~45 vs. ~62 points] and were more likely to have been punished in round 1 [46% vs. 16% punished].)

Participants could pay 0-50 points (in 10-point increments) to punish their partner's response, whether the partner had defected or reciprocated. The mean of amount of punishment inflicted was 8.15 points (*SD* 15.61; range 0-50; median = 0). As expected, the mean was higher in response to defection than reciprocation: 13.59 (*SD* 18.63) vs. 3.19 (*SD* 9.91), t (738.69) = 11.09, $p = 10^{-16}$).

When a responder defects, trusters who risked more suffer greater losses; many theories predict that greater losses will up-regulate motivations to impose negative sanctions—whether these involve withdrawing benefits or inflicting costs (Petersen et al., 2012). *Trust*—the number of points the truster risked—was indeed correlated with the desire to withdraw benefits by leaving a defecting partner: r (242) = .31 ($p = 10^{-6}$). Inflicting costs showed the same pattern and effect size: Participants who sent more points as truster paid more to punish a defecting partner: r (494) = .31 ($p = 10^{-12}$). For this reason, the analyses that follow control for both *Trust*, that is, P, the number of points the participant risked, and defection by the sham responder.

2.3.2.2 Did participants punish less when they thought others can exercise partner choice?

The results on partner switching (section 2.3.1) showed that being punished drives partners away: Inflicting punishment decreases the probability that others will choose to

partner with you. This is a liability when other people can exercise partner choice. We therefore predicted that the perception that others can easily switch partners will down-regulate motivations to punish. To test this prediction, we assessed the amount of punishment delivered as a function of relational mobility—an estimate of partner choice in one's local social ecology—and verbal cues about partner choice in the immediate situation, which were delivered with the instructions. We also analytically controlled for society: US versus Japan.

To determine which of these variables predicted amount of punishment delivered (*Amount paid to punish the responder*: 0-50), we conducted multiple linear regression with the glm function in R (R Core Team, 2020). The predictors examined were condition (High Partner Choice = 1, Low Partner Choice = 0), society (US = 1, Japan = 0), and participants' perceived relational mobility (*RM others* and *RM self*). These analyses controlled for whether the responder reciprocated or defected on the participant (*Defection by the responder*: 50% returned = 0, 20% returned = 1) and how many points participants entrusted to their partner (*Trust*: 0-100).

We also entered interactions between the predictors and, with stepwise selection, determined the best model (using the step function in R (R Core Team, 2020)). Based on AIC scores, the interactions and *RM self* were removed from the model. All continuous variables were centered by subtracting the mean to avoid multicollinearity issues (Robinson & Schumacker, 2009) (resulting VIF values < 1.3).

In determining responses, directly experiencing how a specific partner behaves should have greater weight than any social ecological variable. We did indeed find an order effect of whether the participant was truster or responder in round 1. There were also huge carry-

over effects of the partner's behavior in round 1 on participants' responses in round 2 (see section 2.3.4). To see whether social ecology variables predict punishment *in the absence of a prior history* with the current partner, we analyzed *Amount paid to punish the responder* by those who played the truster role first (n = 509). (Of these, n = 238 experienced a responder who defected.) See tables 1.3a and 1.3b for full models with unstandardized coefficients, associated confidence intervals, and adjusted R^2.

Table 1.3a. Factors affecting the amount paid to punish the responder.

Predictors	*b*	*SE*	95% CI	*B*	*t*	*p*
Condition: High Partner Choice (vs. Low)	-0.06	1.18	[-2.39, 2.27]	-0.002	-0.05	.961
Society: US (vs. Japan)	3.82	1.30	[1.27, 6.37]	0.13	2.94	.003
RM others	-2.89	0.92	[-4.71, -1.08]	-0.14	-3.13	.002
Defection by the responder (1, 0)	8.49	1.19	[6.15, 10.82]	0.30	7.15	< .001
Trust (0-100)	0.06	0.02	[0.02, 0.10]	0.14	3.34	< .001

Note. Adjusted R^2 = 0.12. CI = confidence interval for *b*. VIF values were < 1.3.

Table 1.3b. Factors affecting the amount paid to punish the responder who had defected.

Predictors	*b*	*SE*	95% CI	*B*	*t*	*p*
Condition: High Partner Choice (vs. Low)	-0.86	2.07	[-4.94, 3.23]	-0.03	-0.41	.679
Society: US (vs. Japan)	3.65	2.25	[-0.79, 8.09]	0.11	1.62	.107
RM others	-4.79	1.63	[-8.00, -1.59]	-0.20	-2.95	.004
Trust (0-100)	0.16	0.03	[0.10, 0.22]	0.31	5.00	< .001

Note. Adjusted R^2 = 0.10. CI = confidence interval for *b*. VIF values were < 1.3.

2.3.2.3 Did telling people they could switch partners decrease their motivation to punish?

No. There was no effect of Low vs. High Partner Choice condition on how much participants paid to punish (β = -0.002, p = .96; n = 509). Telling them in advance whether they will play all rounds with their current partner (Low Partner Choice) or have the opportunity to switch partners after round 2 (High Partner Choice condition) did not

influence their punishment decisions, even if we restrict the analysis to the participants who experienced defection (β = -0.03, p = .68; n = 238). Note that this is not because participants were insensitive to the partner choice condition: We found that the condition did affect *Reciprocation by the participant* (DV1) (see section 2.3.3.1).

2.3.2.4 Did how much people punished differ by society (US vs. Japan)?

Yes. All else equal, American participants paid more to punish their responder than Japanese participants did (β = 0.13, p = .003, n = 487). The effect of society was similar (but not significant) when the analysis is restricted to the 238 participants whose responder defected (β = 0.11, p = .107).

2.3.2.5 Did perceptions of relational mobility in their local social ecology affect how much people paid to punish?

Yes: The higher their *RM others* score, the less participants paid to punish their partners (β = -0.14, p = .002, n = 487; when analyzing only those whose responder defected, β = -0.20, p = .004, n = 238). That is, the more opportunities they think others have to form new relationships, the less participants punished their partners. Equivalently: Those who assume the average person in their social ecology has fewer outside options inflicted more punishment. Figure 1.4 illustrates that this negative association holds, regardless of condition and society.

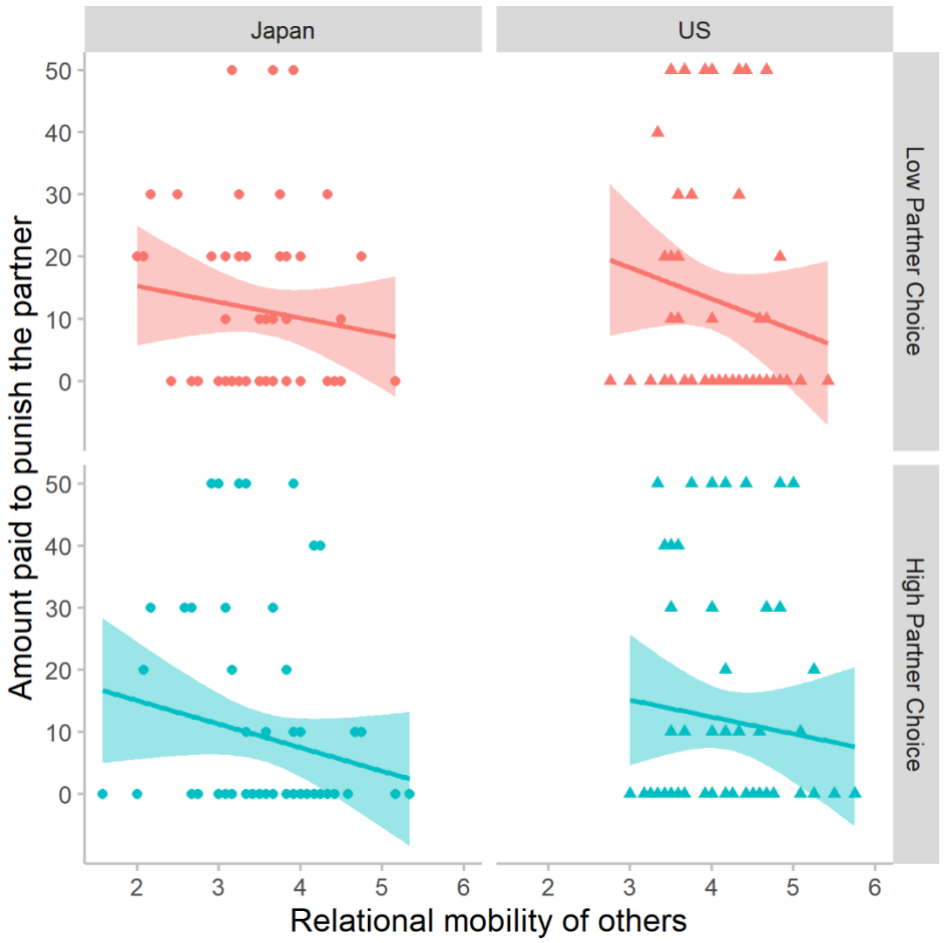


Figure 1.4. The effect of perceived relational mobility of others on punishment. Perceptions of other people's relational mobility (*RM others*) was negatively associated with how much the participant paid to punish their partner who defected. The more they thought others could exercise partner choice, the less they punished the partner.

2.3.3 What predicts reciprocation?

Reciprocation by the participant refers to the percent of 3P points that the participant returned to the (sham) truster (3P = 210). The mean returned was 43.72% of 3P (*SD* = 21.43) and the median was 50%; that is, most responders gave their partner a positive payoff by returning 40% to 100%. For those who responded in the first round (n = 530)—before

experiencing any punishment—71% gave their partner a positive payoff (67% of Americans, 76% of Japanese).

The more partner choice other people can exercise, the more motivated one should be to be seen as a good cooperative partner. This led us to predict that motivations to reciprocate would be up-regulated by perceptions that other people can easily switch partners. To test this prediction, we conducted a multiple regression for *Reciprocation* in the same way as for *Amount paid to punish the responder* (predictors: condition, society, *RM others*, *RM self*, and their interaction terms). *RM self* was removed from the model based on AIC scores. After centering the continuous variables (see section 2.3.2.2), we found no evidence of multicollinearity (VIF values < 4.4). As before, we only analyzed *Reciprocation* by participants who played the responder role first, to avoid carry-over effects from round 1 (*n* = 530; 270 Americans and 260 Japanese). See tables 1.4a, 1.4b, 1.4c for full models.

Table 1.4a. Factors affecting reciprocation by the participant.

Predictors	*b*	*SE*	95% CI	*β*	*t*	*p*
Condition: High Partner Choice (vs. Low)	4.43	2.82	[-1.12, 9.97]	0.10	1.57	.118
Society: US (vs. Japan)	-5.68	2.71	[-10.99, -0.36]	-0.13	-2.10	.036
RM others	8.38	2.49	[3.49, 13.27]	0.27	3.37	< .001
Condition × Society	-9.00	3.96	[-16.78, -1.22]	-0.18	-2.27	.023
Condition × *RM others*	-5.74	3.87	[-13.35, 1.87]	-0.13	-1.48	.139
Society × *RM others*	-11.41	3.93	[-19.14, -3.69]	-0.24	-2.90	.004
Condition × Society × *RM others*	15.80	5.70	[4.61, 26.99]	0.24	2.77	.006

Note. Adjusted $R^2 = 0.07$. CI = confidence interval for *b*. VIF values were < 4.4.

Table 1.4b. Factors affecting reciprocation by Japanese participants.

Predictors	*b*	*SE*	95% CI	*β*	*t*	*p*
Condition: High Partner Choice (vs. Low)	4.43	3.07	[-1.63, 10.48]	0.10	1.44	.151
RM others	8.38	2.71	[3.04, 13.72]	0.25	3.09	.002
Condition × *RM others*	-5.74	4.22	[-14.05, 2.56]	-0.11	-1.36	.175

Note. Adjusted $R^2 = 0.04$. CI = confidence interval for *b*. VIF values were < 2.0.

Table 1.4c. Factors affecting reciprocation by American participants.

Predictors	*b*	*SE*	95% CI	*B*	*t*	*p*
Condition: High Partner Choice (vs. Low)	-4.57	2.52	[-9.53, 0.38]	-0.12	-1.82	.070
RM others	-3.03	2.76	[-8.46, 2.40]	-0.10	-1.10	.273
Condition × *RM others*	10.05	3.78	[2.60, 17.51]	0.24	2.66	.008

Note. Adjusted $R^2 = 0.02$. CI = confidence interval for *b*. VIF values were < 2.3.

2.3.3.1 Did telling people they could switch partners increase reciprocation?

Even though we found no effect of condition on participants' punishment behaviors, it did significantly influence their motivation to reciprocate. Telling participants in advance whether they would play all rounds with their current partner (Low partner choice) or have the opportunity to switch partners after round 2 (High Partner Choice condition) had no main effect on *Reciprocation by the participant* ($\beta = 0.10$, $p = .118$), but it interacted with the other predictors.

There was a 2-way interaction between condition and society ($\beta = -0.18$, $p = .023$) and a 3-way interaction between condition, society, and *RM others* ($\beta = 0.24$, $p = .006$). These significant interaction effects indicate that participants did detect, register, and respond to the verbal cue about the possibility of partner choice in their immediate situation. To examine these interactions, which all involve society, we ran the same regression model for each society separately (see section 2.3.3.4).

2.3.3.2 Did society (US vs. Japan) affect reciprocation?

Yes. Japanese participants returned a larger percentage of the 3P points sent by the (sham) truster than Americans did ($\beta = -0.13$, $p = .036$). The difference between societies

was carried by the extremes. Americans were more likely than Japanese to defect by returning 30% or less (33% vs. 24%, $Z = 2.64$, $p = .008$ [90/270 vs. 63/260]) and less likely to reciprocate generously by returning 60% or more (13% vs. 30%, $Z = -4.74$, $p = 10^{-5}$ [36/270 vs. 78/260]). There was no difference in how likely American and Japanese participants were to return 40% (16% vs. 15% [43/270 vs. 40/260]) or 50% (38% vs. 30% [101/270 vs. 79/260]). There was also a 2-way interaction between society and *RM others* ($\beta = -0.24$, $p = .004$) (see section 2.3.3.4).

2.3.3.3 Did perceptions of relational mobility in their local social ecology affect participants' motivations to reciprocate?

Yes. Participants' motivation to reciprocate their partner's trust was up-regulated by their perceptions of relational mobility in their society (*RM others*: $\beta = 0.27$, $p = .0008$). The more opportunities they thought people in their social ecology have to leave unsatisfying relationships for better ones, the larger the percentage of 3P points they returned as responders (i.e., those who thought others have fewer opportunities to change relationship partners reciprocated less). Figure 1.5 illustrates that this positive association generally holds, regardless of society and condition, except for American participants in the Low Partner Choice condition.

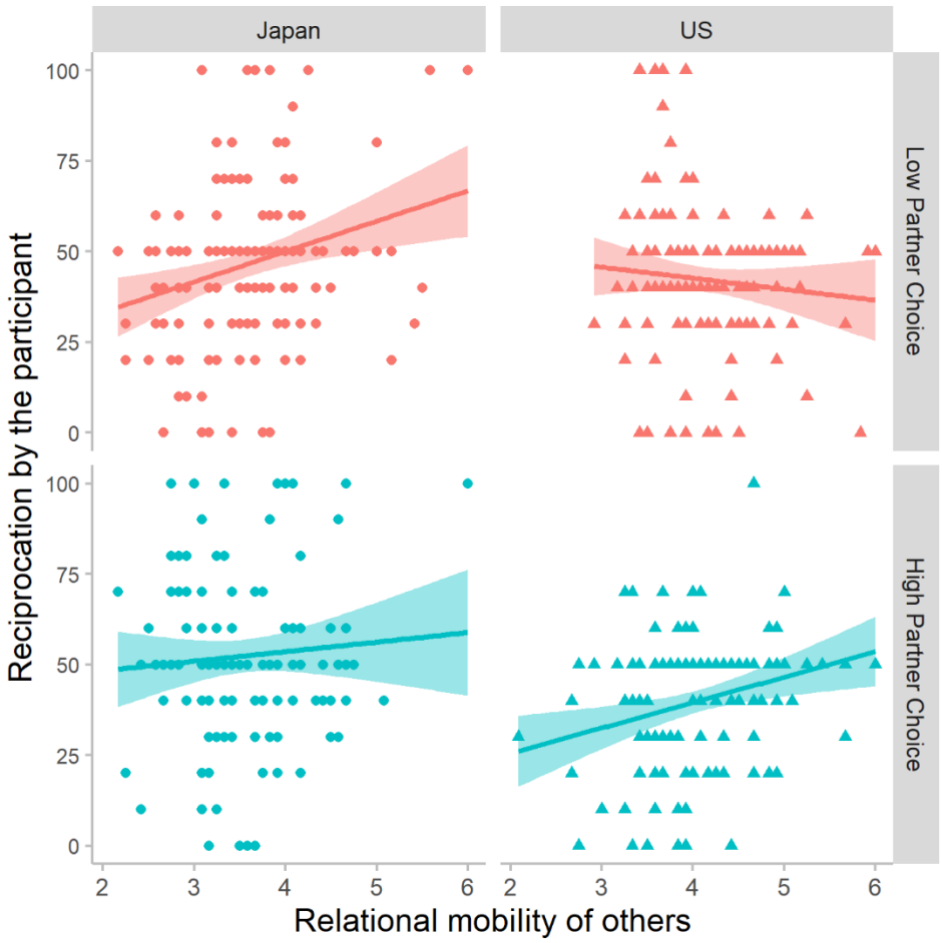


Figure 1.5. The effect of perceived relational mobility of others on reciprocation. Perceptions of other people's relational mobility (*RM others*) was positively associated with the percentage of points the participant returned to the partner (0-100%). The more they thought others could exercise partner choice, the more they reciprocated.

2.3.3.4 How did condition and perceived relational mobility interact in each society?

The patterns shown in figure 1.5 suggest that the effect of perceived relational mobility might differ across conditions and societies. Indeed, there was a 3-way interaction between condition, society, and *RM others* (see section 2.3.3.1). To examine this 3-way interaction, below we analyze the interaction between condition and *RM others* separately in Japan and the US.

Japanese participants (n = 260) up-regulated their motivation to reciprocate with their estimate of other people's relational mobility (*RM others*: β = 0.25, p = .002); the effect size is about the same as when both societies were analyzed together. In Japan there was no significant main effect of condition (β = 0.10, p = .15) or interaction between condition and *RM others* (β = -0.11, p =. 175).

For American participants (n = 270), there was no main effect of *RM others* (β = -0.10, p = .273), but there was an interaction (β = 0.24, p = .008) between this variable and condition—whether they were told that they would play all rounds with the same partner or have the opportunity to switch after round 2. When told they would have the opportunity to switch partners, American's perceptions of relational mobility regulated reciprocation, with the same effect size as found in Japan (High Partner Choice condition, *RM others* predicting *Reciprocation by the participant*: β = 0.25, p = .004, n = 128). This relationship was absent when Americans were told they would always interact with the same partner (Low Partner Choice condition, *RM others* predicting *Reciprocation by the participant*: β = -0.09, p = .306, n = 142). This is not because Americans treated their partners poorly when the situation precluded partner choice: Overall levels of reciprocation were similar across both conditions—42% (Low) vs. 40% (High)—even when controlling for *RM others* (β = -0.12, p = .07).

If relational mobility represents the prior probability that people can find new relationship partners in their local social ecology, then Americans updated that prior based on verbal cues regarding the immediate situation, but Japanese participants did not.

2.3.4 Are qualities of the person used to update priors based on social ecology?

Priors based on social ecology are most relevant when you have no other information about a new partner. This prior should be most strongly updated by learning what your new partner is like (Jussim, 1991), with their actual behavior toward you a good cue to how they treat strangers. All else equal, a good cooperator will have more outside options than a defector, so efforts to retain your partner should increase with evidence that this partner is a valuable cooperator.

Consistent with this view, the partner's behavior in round 1 influenced participants' behavior in round 2, with little or no remaining effect of social ecology. For example, reciprocation in round 2—the percent of 210 points that participants returned—reflected reciprocation vs. defection by the sham partner in round 1: They returned much more to sham partners who had reciprocated instead of defecting in round 1 (47.1% vs. 36.8%, $t(457.72) = 5.49$, $p = 10^{-8}$). Controlling for all other factors, defection by the sham partner in round 1 predicted participants' reciprocation in round 2 ($\beta = -.23$, $p = 10^{-7}$), but their perceptions of relational mobility did not (controlling for round 1 behaviors, *RM others*: $\beta = .06$, $p = .345$). Similarly, when the partner defected in round 2, relational mobility had no influence on whether participants punished them ($\beta = -.11$, $p = .092$, controlling for other factors).

Participants who returned 40% or less in round 1 had a 50% chance of being punished. When these participants were trusters in round 2, relational mobility had no effect on how many points they risked, but their partner's round 1 behavior did. Participants who had not been punished in round 1 risked more points in round 2 than those who had been punished

(53.60 vs. 40.16, t (230.91) = 3.13, p =.002; controlling for other factors, *Punishment received* in round 1: β = -.22, p =.0002). This effect was strongest in those who were punished for returning 40% in round 1 (β = -.41, p =.0001). Those who were punished in round 1 also paid more to punish in round 2 than those who were not (12.05 vs. 6.49 points, t (233.79) = 2.76, p = .006; controlling for other factors, β = .23, p =.0001). That is, many engaged in retaliatory punishment.

In round 2, higher relational mobility had just one effect: Americans who believe people in their social ecology can easily switch partners were less likely to punish a partner who had reciprocated their trust (i.e., it decreased anti-social punishment in round 2; β = -.31, p =.012, controlling for condition).

2.4. Study 1 Discussion

2.4.1 Evidence that motivational systems are designed for social ecologies with varying levels of partner choice

Ancestral variation in the availability of cooperative partners would have favored the evolution of motivational systems that treat partner choice as a continuous variable. Motivations to keep valuable cooperative partners and abandon unrewarding ones should be up-regulated in response to the perception that others can easily switch partners.

Here we tested the hypothesis that an individual's motivations to reciprocate and punish are calibrated by that person's estimate of the degree to which others in their local social ecology can exercise partner choice. This estimate is captured by measures of *relational mobility* (Yuki et al., 2007). The higher an individual's relational mobility score, the more opportunities they believe others have to leave unsatisfying relationships for better ones.

We assessed motivations to trust, reciprocate, defect, punish, and switch partners by allowing people to cooperate for mutual benefit with a new individual. The results showed that motivations to reciprocate and punish tracked participants' perceptions of relational mobility. The more partner choice they thought others in their social ecology could exercise, the more they reciprocated their partner's trust and the less they paid to punish their partner—even when that partner had defected.

Providing incentives for desirable partners to stay in the relationship is the proposed function of these motivational calibrations. If that is correct, then people who have the opportunity to switch partners will be more likely to stay with a partner who reciprocates their trust and more likely to leave one who punishes them. After two rounds, half the participants were asked if they wanted to keep their current partner or switch to someone new. Holding all else equal, having been defected on more than quadrupled the odds that they wanted to switch and having been punished tripled the odds they would choose to leave. These were the two biggest independent predictors of switching decisions. The desire to leave a partner who punished was especially strong for participants who returned 40%—a response that creates a positive payoff for both parties that is almost equal. These individuals were almost 10 times more likely to want a new partner.

2.4.1.1 Are priors about social ecology updated by information about the situation or the person?

Perceptions of relational mobility are based on a huge database of experiences in a local social ecology—sometimes a lifetime's worth. For this reason, we proposed that relational mobility serves as an estimate of the prior probability that others in one's social ecology can

exercise partner choice. It is a best guess before you learn what your partner is like—the situation participants faced in round 1.

If relational mobility in your social ecology is used to estimate a partner's outside options when you know nothing else about that person, then its effect on cooperative motivations should be reduced (or eliminated) by data about that specific person's value as a cooperative partner—to yourself and others. The evidence indicates that participants in both societies updated this prior based on first-hand knowledge of their partner's willingness to cooperate and reluctance to punish. Once participants had experienced how their partner behaved in round 1, relational mobility no longer predicted how much they trusted, reciprocated, or punished in round 2, in either the US or Japan. The behavior of the sham partner in round 1 (and, of course, in round 2) did predict their responses. The only behavior that relational mobility continued to influence was antisocial punishment. The belief others in your social ecology can easily switch partners tempered—but did not eliminate—antisocial punishment.

The results suggest that estimates of partner choice based on social ecology are updated based on properties of the *person* with whom one is interacting. But are these estimates updated in response to cues about a temporary *situation* one is facing—ones unrelated to the partner's value as a cooperator? It is not clear that they should be.

Delton and colleagues (2011) examined the evolution of motivations to cooperate in Bayesian agents who knew the base rate of one-shot interactions in their population and updated this prior based on a cue about the immediate situation they were facing. The cue reflected the probability that they would never interact again with their current partner. These Bayesian agents evolved a strong disposition to cooperate *even when they rationally*

believed the interaction was one-shot. Selection favored agents who behaved *as if* they would repeatedly interact with their current partner even when they knew this was unlikely. Agent-based models also show that meeting a new individual once was a good cue that you will meet them again in ancestral social ecologies (Krasnow et al., 2013) . Every participant in our study was exposed to this ancestrally-reliable cue to a shadow of the future: They interacted with their partner for two rounds.

We did, however, provide a verbal cue relevant to partner choice in the temporary situation that they were facing. Half the participants were told they would be interacting with the same partner in every round (i.e., they were engaged in a repeated interaction with this person). The other half were told they could change partners after two rounds (i.e., their current partner can refuse to interact with them repeatedly). If this verbal cue is used to (temporarily) update their prior probability that a newly encountered person can exercise partner choice, their motivations to cooperate or punish might shift in response.

There was little evidence that participants in round 1 used this situational cue to update a prior that was based on their social ecology. Being told whether they would have the opportunity to switch partners had no effect on how much participants punished defections by their partner: Higher relational mobility in their local social ecology predicted less punishment, regardless of condition or society. The cue did have an effect on how much American participants reciprocated their partner's trust, however. Although average levels of reciprocation were similar in both conditions, higher relational mobility predicted more reciprocation when Americans were told they and their partner could part ways after two rounds, but not when they were told that all of their interactions would be with the same partner.

Japanese participants did not respond to this cue at all: Their estimates of relational mobility predicted more reciprocation (and less punishment) to the same extent in both conditions. That is, there was no evidence that people in Japan updated their prior hypothesis about relational mobility based on the situational cue we provided. If they did, the change was too small to influence their willingness to reciprocate or punish.

If this result generalizes to other cues about a temporary situation, it suggests that the benefits of opportunistic behavior in the short term were generally outweighed by the risk of losing a valuable, long-term cooperative partner.

2.4.2 What is the function of punishment in dyadic reciprocal cooperation?

What, if anything, is the adaptive function of motivations to pay a cost to punish a defecting partner? This was not a rare response: Of participants who were trusters in round 1, 44% punished when the responder defected. It is usually assumed that the function of punishing defectors is to elicit more cooperation from them in the future—especially when they do not have the option to change partners.

People who believe others in their social ecology have fewer options to switch partners did pay more to punish defectors: Low relational mobility scores predicted paying more to punish. But there was no evidence that punishment succeeded in eliciting greater cooperation from participants. Quite the contrary: Participants who were punished for returning 0-40% in round 1 did not respond by sending more points as truster in round 2. Indeed, they sent fewer points as truster ($\beta = -.22$, $p = .0002$), and this effect was particularly pronounced for those who had provided a positive payoff by returning 40% in round 1, $\beta =$ -

.41, $p = .0001$ (vs. $\beta = -.12$, $p = .099$ for those who provided a negative payoff in round 1). Moreover, those who were punished in round 1 were more likely to retaliate by punishing their partner in round 2 (for similar results, see Bone et al., 2015, 2016).

Not only did punishment fail to elicit more cooperation from punished partners, but it also drove them away. When partner switching was possible, having been punished was one of the biggest independent predictors of wanting to change partners. Driving away defectors might be a function of punishment, of course—when they were not punished, ~70% of people who returned 0-40% wanted to remain with their accommodating partner (~68% of those who returned 0-30%; ~71% of those returning 40%). Although participants in this study could prevent future interactions at lower cost by simply deciding to switch after round 2, avoiding unrewarding partners may be more difficult in real life, especially when they want to continue cooperating with you.

Krasnow et al. (2012) suggest that punishing defection signals a willingness to continue cooperating with your current partner, but on more favorable terms. Using a paradigm similar to the TGP, they found that participants who punished a defecting partner in the first round were 11 times more likely to cooperate than defect in the second one (switching was not an option). This pattern was not apparent in our study: Participants who punished a defecting partner did not return more in round 2 than those who did not (39.28% vs. 35.18%, t (226.99) = 1.41, $p = .160$), and they were not more likely to want to remain with their partner—indeed, the more points participants paid to punish the partner, the more—although slightly—they wanted to switch (OR = 1.02; 95% CI = [1.01, 1.03]). (Note, however, that a participant's decision to stay did not ensure a continuing interaction in our study; the partner also had the option to leave, and punished ones were likely to do so.)

Our results showing that retaliatory punishment was common—~45% of those who were punished in round 1 retaliated in round 2—suggest an alternative explanation. In Krasnow et al. (2012), participants who punished defectors in round 1 may have cooperated in round 2 to avoid (very costly) retaliatory punishment by their partner. Those who did not punish partners who succumbed to the temptation to cheat in round 1 may have assumed their partner would "reciprocate" by not punishing them when they did the same in round 2.

Motivations to punish did not reflect the participant's own commitment to stay in the relationship, but they were up-regulated by estimates that *partners* might have few outside options: Lower relational mobility in one's social ecology did predict amount paid to punish defectors. The results are consistent with the hypothesis that motivations to punish evolved to deter bad treatment in the future by partners who do not seem to value your welfare (Krasnow et al., 2016). Defecting now may be a reliable cue that this partner does not value your welfare sufficiently, and punishment was overwhelmingly directed at defectors. In ancestral social ecologies, partners who part ways now may nevertheless have to cooperate again in the future (Krasnow et al., 2013, 2016; Smith et al., 2018). Punishment may have evolved as a warning, to deter bad treatment by defectors who may darken your door in the future.

2.4.3 Micro and macro effects of social ecology

We measured two variables regarding participants' real-life social ecology of partner choice. First, we measured participants' perceptions of their partner choice ecology with the relational mobility scale (Yuki et al., 2007). Second, we recruited participants from two societies in which average relational mobility scores are typically high (US) versus low

(Japan). This lets us see whether behavior at the individual level scales up to explain differences between nations.

Within each society, the motivations of individuals were calibrated by their perceptions of other people's relational mobility: the number of opportunities they believe that others have to form new relationships. Moreover, the pattern of calibration was universal: Within each society, higher relational mobility scores predicted more reciprocation and less punishment. Individual-level effects tracked individual perceptions of the local social ecology.

What about group-level differences? The concept of relational mobility was built from Yamagishi's seminal work on general trust: a cognitive bias to assume that newly encountered people will treat you with benevolence rather than exploitation (Schug et al., 2009; Yamagishi, 2011). *General trust* varies across nations; scores on the standard survey measure are higher in the US than Japan, for example. Where general trust is higher, people are more willing to risk cooperating with strangers who could, if untrustworthy, profit at their expense. The benefit of trusting strangers is that it allows people to discover better cooperative partners, giving them more outside options. The resulting increase in relational mobility then tempers the risk of trusting strangers: The threat that a good partner will leave for a better outside option can deter exploitive behavior and increase benevolence.

With this in mind, we compared average behavior in the US and Japan. As in other studies, perceptions of relational mobility were higher in the US than Japan (*RM others*: 4.12 vs. 3.57, t (1028.2) = 13.76, $p = 10^{-16}$; *RM self*: 4.20 vs. 3.37, t (1030.8) = 18.71, $p = 10^{-16}$). That is, the average American believes others have more outside options than the average person from Japan does. Moreover, as Yamagishi's view of general trust predicts,

when participants had no prior experiences with their partners, American trusters risked more points on a stranger than Japanese participants did (*Trust*: 59 vs. 50.6, t (502.55) = 2.9, p = .004). And trusting strangers usually paid off: Most responders delivered a positive payoff in both societies (US 67%, JP 76%).

Did the perception that others have more outside options lead the average American to reciprocate more and punish less than the average person from Japan? No. Not only did Americans return less, on average, than Japanese participants, but more of them exploited their partner's trust by delivering a negative payoff (US 33% vs. JP 24%). Americans were also more punitive, not less: They paid more to punish their partners, even when controlling for all other factors (including whether their partner defected). And, despite less reciprocation and more punishment at the macro-level, Americans were more likely to stay with their partner than Japanese participants (all else equal).

Within each society, individual differences in reciprocation and punishment were associated with individual differences in perceptions of relational mobility, but this did not translate into group-level differences between the US and Japan. Assuming that individual differences fully explain group-level differences is called the *ecological fallacy* (Brewer & Venaik, 2014; Pollet et al., 2014; Thorndike, 1939). The data clearly show that the micro-level effect of individuals' perceptions of relational mobility and the macro-level effect of society were independent of one another. The individual-level psychological calibrations and the group-level differences between nations coexist, rather than one producing the other.

Features of the social ecology other than relational mobility could be responsible for the differences in group-level calibrations between the US and Japan (see e.g., (Hashimoto & Yamagishi, 2016; Yamagishi et al., 2008)). That Japanese participants were less punitive

than Americans is contrary to findings that Japan (or East Asian countries in general) has "tighter" norms than the US which, when broken, elicit great censure (M. J. Gelfand et al., 2011; Wang & Leung, 2010), but perhaps consistent with studies showing greater motivations to avoid rejection in people in Japan than the US (Hashimoto & Yamagishi, 2016). Our data cannot speak to these explanations of the group-level differences we found.

2.4.4 Limitations and future directions

Motivations responded when participants learned how the partner treats them, but the partner switching instructions influenced Americans only (and not much at that). This could be because repeated interactions—with interruptions between—were common ancestrally, making long-run estimates of social ecology a more reliable basis for calibration than cues about a fleeting situation. The other possibility is that a cue delivered online was too divorced from real life, devoid of psychophysical cues typical of social isolation versus community. Future studies might enhance the salience of the situational cue, perhaps by including visual displays showing many versus few alternative partners (avatars or faces), or by giving participants prior experiences of a desirable partner leaving for a better one or an unrewarding partner staying.

A person with fewer outside options than others in their local ecology may feel they need to reciprocate more and punish less. We did adapt the relational mobility scale to ask about the self; although self and other scores were correlated r (515) = .60 ($p = 10^{-16}$) in the US and r (516) = .50 ($p = 10^{-16}$) in Japan, we calculated whether *RM self* < *RM other* for each participant. In Japan, 67% of participants felt their outside options were worse than those of other people, compared to 44% in the US. And, in both countries, those who felt

they have fewer outside options returned more points than those who felt their options were better than or equal to others, but the difference in points returned was not significant. A better measure in the future might be to ask, for each RM question, whether people feel they have more, the same, or fewer options than others in their society.

Dyadic cooperation may be affected by other aspects of the social ecology as well, such as how likely others will be to take advantage of you (Yamagishi, 2011). Punishment as a deterrent may be up-regulated in ecologies where the probability of being exploited are higher, as they were in the US in this study. Perceptions of these probabilities would be a fruitful variable to assess.

Lastly, our participants were from either the US or Japan, two populous, large-scale industrialized societies. Objectively speaking, most people in these countries are free to associate with anyone they like, and they are surrounded by strangers, each of whom is a potential new partner. It would be fruitful to extend the current line of research to smaller societies in which the actual—not only perceived—possibility of partner choice is more limited.

2.4.5 Conclusions

The results of study 1 demonstrate that estimates of partner choice in one's local social ecology regulate motivations to reciprocate, defect, and punish in dyadic cooperative interactions. The more opportunities participants thought others have to form new relationships, the more they reciprocated and the less they punished. The results suggest that the function of these motivational calibrations is to retain valuable cooperative partners when they have the option to leave: When given the choice, participants preferred to stay

with partners who reciprocated and leave partners who punished them. The results support the hypothesis that motivational systems are designed to use estimates of the degree of partner choice in one's local social ecology to functionally regulate reciprocal behaviors.

Chapter 3: Study 2. Motivational regulations based on reputation concern

3.1 Study 2 Introduction

Study 1 demonstrated that estimates of how easily others can find a new partner—a cue of competition to be chosen as a cooperation partner (Baumard et al., 2013; Debove et al., 2015)—regulate motivations to cooperate and punish in dyadic social exchange. However, the presence of competition alone does not mean *you* are in a position to compete to be chosen. To enter competition and be chosen, you first need to be recognized as a potential partner and then invest in your reputation as a cooperator. Study 2 tests how the possibility of being recognized and evaluated as a potential partner affects how motivations are regulated in dyadic social exchange.

3.1.1 Motivations to cooperate when you are being evaluated as a partner

Cooperating and behaving generously in the presence of others is the most straightforward way to invest in your reputation as a valuable cooperation partner (Barclay, 2013; 2016). Yet, an effect of a generous behavior will differ by situations (e.g., donating money when nobody is watching versus others are watching). Systems regulating motivations to cooperate should register these situational cues as inputs and calibrate behaviors accordingly. Indeed, a meta-analysis shows that there is a small but significant increase in generosity when people believe that they are being observed (Bradley et al.,

2018). However, many studies show that people act more generously when they believe that others are not only observing but actively evaluating them (Sommerfeld et al., 2007; Boero et al., 2009; Feinberg et al., 2014; J. Wu et al., 2015, 2016b, 2016a; Sommerfeld et al., 2008). And notably, people behave most cooperatively when observers exert partner choice based on their observations (Barclay, 2004; Barclay & Willer, 2007; Sylwester & Roberts, 2010, 2013). These findings suggest that the key input to motivational systems is not the presence of observers per se but whether one is being recognized and assessed as a cooperation partner.

3.1.1.1 Group as a cue of reputation-based partner choice

Several theories suggest that group is a cue that one is being viewed and evaluated as a potential cooperation partner. Yamagishi and colleagues argue that group situations elicit concern for managing one's reputation as a cooperator (Mifune et al., 2010; Yamagishi & Mifune, 2008, 2009, 2016). They propose that, because most social exchanges take place within demarcated groups, the presence of a salient ingroup activates a "group heuristic" in the mind. According to the group heuristic hypothesis, people (i) assume that there are generalized reciprocal exchanges in a group and expect favorable treatment from members of one's own group (ingroup). People therefore (ii) behave in a way that minimizes the risk of developing a bad reputation for being a bad cooperator—a free rider or cheater—among the ingroup members. Here, cooperation with a member of one's own group is a "default" decision strategy to safeguard one's reputation as a cooperator and prevent exclusion from social exchange.

The group heuristic account was originally offered to explain ingroup favoritism: a tendency to treat members of your own group more favorably than members of different groups.[4] People favor ingroup members even in the "minimal group" setting, an experimental situation where participants are categorized into artificial groups based on trivial features (e.g., preference for paintings by Kandinsky vs. Klee) without any communication or interaction between group members (Tajfel et al., 1971).

Reviewing decades of research on ingroup favoritism, Pietraszewski (2013, 2021) points out that group is a marker of an opportunity to form a cooperative relationship and thus group membership is a potent cue of reputation-based partner choice. Pietraszewski proposes that humans have a cognitive tool kit to navigate the complex social world: coalitional psychology (Pietraszewski, 2013, 2021). It (i) detects and keeps track of patterns of social relationships such as coalition and cooperation, and competition, (ii) predicts behaviors based on the social relationships, and (iii) motivates one to initiate relationships based on the former two functions. According to this account, when detecting cues suggesting someone is a member of one's own group (function i), the mind would perceive it as an opportunity to probe for cooperative intent and establish a cooperative relationship (function iii), predicting that the ingroup member would be also interested in initiating

[4] Yamagishi and colleagues originally argued that the expectations of favorable treatment from ingroup members produce ingroup-favoring cooperation (Jin & Yamagishi, 1997; Yamagishi et al., 1999; Yamagishi & Kiyonari, 2000). They submitted this updated reputational account after the finding that people show ingroup-favoring cooperation even when they cannot expect more reciprocation from an ingroup member than from an outgroup member (e.g., when playing the second player in a sequential Prisoner's Dilemma game where the first player, either an ingroup or outgroup member, already decided to cooperate (Horita & Yamagishi, 2010; Simpson, 2006); see Yamagishi & Mifune (2008) for details). The two versions nonetheless provide the same ultimate explanation for ingroup-favoring cooperation, only differing in the proximate psychological explanation (expectations vs. reputation management) that they shed light on.

cooperation (function ii). In other words, the mind regards one's own group as a pool of potential partners, expecting that members of one's own group also consider one to be a potential partner and will therefore assess one's reputations as a cooperator.

3.1.1.2 Alternative hypothesis: ingroup favoritism based on social identity processes

Ingroup favoritism is commonly attributed to social identity processes (Dunham, 2018; Everett et al., 2015a; Spears, 2021). Social identity theory (Tajfel, 1982) posits that (i) people identify with their own social group and have a self-concept based on their membership of the group ("social identity"). People (ii) are therefore motivated to positively differentiate their own group from different groups (outgroups)—e.g., allocate more monetary rewards to ingroup than outgroup members—to enhance their social identity. These psychological processes behind favoring one's own group for the sake of the group are also referred to as parochialism (parochial altruism) (Everett et al., 2015a; Yamagishi & Mifune, 2016) or ethnocentrism (Tajfel, 1982).

These explanations based on theories of social identity argue that ingroup favoritism is driven by preferences for your own group and motivations to increase the payoffs of your fellow group members. Here, behaving cooperatively toward ingroup members is irrelevant to a concern with managing your own reputational with ingroup members as potential partners. Therefore, these explanations predict ingroup favoritism regardless of whether others know you are behaving this way or not—the concern for reputation should not alter your motivations to positively differentiate your own group from other groups.

3.1.1.3 Eliminating ingroup favoritism by manipulating reputation concern

However, a series of experiments by Yamagishi and colleagues demonstrates that a simple manipulation can eliminate ingroup favoritism in the minimal group setting by erasing the concern for reputation management (Jin & Yamagishi, 1997; Mifune et al., 2010; Yamagishi et al., 1999; Yamagishi & Mifune, 2008).

Their experiments had a common set-up as follows. In one condition, participants could give money to an ingroup member; in another condition, the recipient was an outgroup. These group membership conditions were crossed with another manipulation: the knowledge of the two players about their group membership. There were two conditions. (i) In the "common knowledge" condition, both the participant and the recipient knew which group they belonged to. (ii) In the "private knowledge" condition, only the participant knew which group they both belonged to, and the recipient did not know which group the participant belonged to. Yamagishi and colleagues predicted that participants would give more money to an ingroup member than to an outgroup member only in the common knowledge condition. According to their reasoning, this is because what elicits reputation concern is being identified as an ingroup member by other ingroup members. When participants knew that they were not identified by their own group members, there is no point to treating ingroup members favorably to avoid obtaining a bad reputation—that is, there is no concern for reputation management.

They found what they predicted: Ingroup-favoring cooperation disappears—or at least decreases—when group membership is not common knowledge (Foddy et al., 2009; Horita & Yamagishi, 2007; Jin & Yamagishi, 1997; Mifune et al., 2010; Yamagishi et al., 1999, 2005; Yamagishi & Mifune, 2008). When participants knew that their ingroup partners were

unaware of who they were, they did not treat the oblivious ingroup partners more favorably than outgroup partners. The result of these experiments illustrates that group membership is a cue of reputation-based partner choice—one is being evaluated as a cooperation partner and thus one's reputation is at stake to be chosen or not.

In summary, evidence suggests that interactions with members of one's own group indicate that one is being considered and evaluated as a cooperation partner. Therefore, all else being equal, the presence of an ingroup member should up-regulate motivations to behave cooperatively; this is a "default" strategy to be chosen as a partner by acquiring a reputation as a cooperator (or avoiding a reputation as a cheater). However, if there is another cue that one is not being evaluated as a partner—such as when an ingroup member is unaware of your group membership—there is no concern for managing one's reputations and thus up-regulating motivations to cooperate should not be observed. This prediction was supported by a number of experiments by Yamagishi and colleagues (e.g., Yamagishi et al., 1999; Yamagishi & Mifune, 2008) and by others using a similar manipulation of group membership knowledge (Guala, 2012; Ockenfels & Werner, 2014; Romano, Balliet, Yamagishi, et al., 2017). A meta-analysis also confirms that the manipulation has a robust effect on ingroup-favoring cooperation (Balliet et al., 2014).

However, recent studies report mixed results. When researchers employ alternative manipulations to control for whether participants are being evaluated by potential partners (e.g., participants were told that their decisions would not be revealed to others or they would not interact with ingroup members), the manipulation sometimes decreased cooperation toward ingroup members (Everett et al., 2015b; Imada, 2020; Misch et al., 2021; Romano, Balliet, & Wu, 2017) but sometimes had no effect at all (Imada, 2020;

Misch et al., 2021). Various factors may explain these conflicting results (Imada, 2020), but the most notable is that the majority of these studies is underpowered (sample sizes up to 121, except for Romano et al., 2017) to detect the small-to-medium sized effect of ingroup-biased cooperation or knowledge manipulation (both $d = 0.32$) (Balliet et al., 2014).

The present study aims to replicate and extend the experiments by Yamagishi and colleagues (Jin & Yamagishi, 1997; Mifune et al., 2010; Yamagishi et al., 1999; Yamagishi & Mifune, 2008) with larger samples (over 200) and using a different manipulation. Like in the experimental set-up in Yamagishi et al., *Group membership* (*ingroup* vs. *outgroup* partner conditions) was crossed with a simple manipulation that is designed to eliminate reputation concern: *Identifiability.* In one condition, participants were identifiable and trackable as an individual because their identities were revealed to others (*identified* condition); in another, participants' identities were hidden from others (*anonymous* condition). The logic of this manipulation is the same as the one in Yamagishi et al.: Even when you are interacting with someone from your own group, if the person does not know who you are, reputation concern is absent—it is impossible for the person to form or keep track of your reputation. If group membership indicates that you are being evaluated as a partner, your motivations to cooperate will be up-regulated during an interaction with an ingroup member rather than with an outgroup member. However, the up-regulation will be suppressed when there is evidence that you cannot be evaluated as a partner—that is, when you are anonymous.

3.1.2 Motivations to punish when you are being evaluated as a partner?

Another goal of the present study is to examine how the possibility of being evaluated as a cooperation partner affects motivations to inflict punishment. Above, I argued that inflicting punishment can harm your reputation for being cooperative and decrease the probability that others agree to partner with you, based on the available evidence (Dhaliwal et al., 2021; Horita, 2010; Kiyonari & Barclay, 2008; Mifune et al., 2020; Ozono & Watabe, 2012). A result of study 1 also provided tentative evidence: Inflicting punishment on a partner drastically increased the probability of the partner leaving the relationship. Therefore, if group membership is a cue of being evaluated as a cooperation partner, one can predict that motivations to inflict punishment will be down-regulated in front of members of your own group.

However, past work suggests that, unlike mechanisms regulating motivations to cooperate, mechanisms regulating punitive motivations may not take group as a simple cue of reputation-based partner choice. Although there are many experiments examining how group membership influences motivations to punish, the results are mixed. Some studies show that ingroup favoritism occurs in the domain of inflicting punishment: People inflict less punishment on cheaters in their own group than on cheaters in a different group (Bernhard, Fehr, et al., 2006; Bernhard, Fischbacher, et al., 2006; Delton & Krasnow, 2017; Goette et al., 2012; Guo et al., 2020; Jordan et al., 2014; Martin et al., 2020; Schiller et al., 2014; Valenzuela & Srivastava, 2012; Yudkin et al., 2016). But some studies provide the opposite pattern: People punish ingroup cheaters more harshly than outgroup cheaters (Mendoza et al., 2014; Shinada et al., 2004).

There are several variables that might account for the mixed evidence (Martin et al., 2020). Lack of reputation concern is one: In the majority of studies suggesting ingroup favoritism in punishment, it was inflicted in third party punishment games (Bernhard, Fehr, et al., 2006; Bernhard, Fischbacher, et al., 2006; Delton & Krasnow, 2017; Goette et al., 2012; Guo et al., 2020; Jordan et al., 2014; Schiller et al., 2014; Yudkin et al., 2016); this was not the case in the studies showing the opposite pattern (Mendoza et al., 2014; Shinada et al., 2004). In these games, a third-party punisher was given no opportunity to interact with an outgroup cheater, neither before nor after their decision to punish. The punisher therefore had no obvious reason to worry about negative consequences of inflicting punishment on the outgroup cheater—it could not hurt their cooperative reputation and then lower the probability of being chosen as a partner. In contrast, when the cheater was a member of the punisher's group, the shared membership could elicit reputation concern by indicating the possibility of being evaluated as a partner (e.g., Pietraszewski, 2021; Yamagishi & Mifune, 2008). That is, third-party punishers may have refrained from punishing ingroup cheaters due to concern for losing their reputation as a cooperator. However, very few studies investigated motivations to punish in group situations as a part of reputation management strategy (for an exception, see Delton & Krasnow, 2017).

The present study tests whether the two cues indicating the possibility of being evaluated as a partner—*Group membership* and *Identifiability*—will elicit reputation concern and thereby down-regulate motivations to punish. If group serves as a cue of reputation-based partner choice to systems regulating motivations to punish, it is predicted that motivations to punish will be down-regulated during an interaction with an ingroup member rather than

with an outgroup member. However, the down-regulation will be attenuated when inflicting punishment is unlikely to harm one's reputation—when one is anonymous.

3.1.3 The current experiments

Study 2 used a less complex economic game than study 1 to measure motivations to cooperate with and punish partners (see section 3.2.1.2.1 below). During an interaction with partners, two cues were presented: (i) *Group membership* of the partner (*ingroup* vs. *outgroup* partner), as an indicator of whether the partner is likely to consider one as a potential partner, and (ii) *Identifiability* of the participant (*identified* vs. *anonymous*), indicating whether the partner is able to form and keep track of the participant's reputations.

Three experiments were conducted with differing situations and samples. Study 2a was conducted with an online sample recruited via a crowd-sourcing website, Prolific. To increase the saliency of the *Identifiability* cue, studies 2b and 2c allowed participants to choose a partner and was conducted with college samples.

3.2 Study 2a: Prolific sample, online

3.2.1 Study 2a Methods

3.2.1.1 Participants

The study was conducted online. Participants were 241 English speakers in the United States (65% female, $M_{age} = 29$, $SD_{age} = 9$) recruited via Prolific. They received 3.17 dollars for their participation, which lasted about 20 minutes.

This study as well as studies 2b and 2c were approved by the Institutional Review Board at University of California, Santa Barbara (Human Subjects Committee). Those who wished to participate in the study first completed an informed consent form. After the study, participants received a written debriefing about the study design and purposes. They were then asked for consent to use their data; it was explained that they would be compensated regardless of their answer. One participant who did not provide consent was excluded from the analysis.

3.2.1.2 Design

The experiment was a 2 × 2 within-subjects design: (i) *Group membership* of the partner (*ingroup* vs. *outgroup* partner) was crossed with (ii) *Identifiability* of the participant (*identified* vs. *anonymous*).

Participants interacted with several different partners in a Dictator Game with Punishment (DGP) (see below). Participants interacted with each partner once. In some DGP rounds, the partner was a member of participant's group (*ingroup*), and in some, the partner was a member of the other group (*outgroup*). The arbitrary groups were created via the minimal group paradigm (Tajfel et al., 1971): Participants were classified into two minimal groups—Team Red or Team Blue—based on which of two words they first found in a word search task ("owl" or "cat") (Martin et al., 2020) (the team-color/word combination was counter-balanced; see Appendix B).

In the beginning of the study, participants were instructed to enter their initials (the first letters of their first and last names). In some DGP rounds, both the participant and the partner saw each other's initials (e.g., "S.A.", "C.P.") throughout the round. In this

condition, participants knew that they could be *identified* by their partners. But in some other rounds, participants were informed that their partner did not see their initials—the participant's initials were shown as "?.?."—while participants could see their partner's initials. In this condition, participants knew that they were *anonymous* to their partners. (See section 3.2.1.2.2 below for details.)

3.2.1.2.1 Cooperation and punishment in the DGP

To measure motivations to cooperate with and punish partners, study 2 used a Dictator Game with Punishment (DGP), a behavioral economic game similar to but simpler than the TGP in study 1. The DGP provides an individual, a giver, with an opportunity to share a benefit with another individual, a receiver.

The DGP had a structure as follows. The giver starts with an endowment of 150 points. The giver decides how many points they would like to share with the partner, the receiver, from 0 to 150 points. The points given to the receiver is the dependent variable that measures the participant's motivation to cooperate (DV1: *Giving*; 0-150 in 10-points increments). After seeing how many points the giver shared, the receiver is given an opportunity to punish the giver. The giver could either (i) pay 10 points to deduct 50 points from the giver or (ii) pay 0 points and do nothing. This binary choice of the receiver to inflict punishment on the giver is the dependent variable measuring the participant's motivation to punish (DV2: *Punishment*; 1 = punish, 0 = not punish). Terms such as "cooperation" or "punishment" were not used in the instructions to participants (see Appendix B).

Before the DGP, participants were told that they were going to be given points that could be used during the interaction (as in study 1). They were asked to imagine that the points

they earned would be converted to real money at the end of the study. Before every round of the DGP, each participant (giver and receiver) was given a bonus of 50 points (see Appendix B). This was to ensure that the receiver had enough points to punish the giver regardless of how many points the giver gave the receiver.

As in study 1, participants were told that they would interact with another participant, but in reality, they interacted with sham partners simulated by a program. This procedure, the only deception in the study, was necessary to examine hypotheses about how people react to various partners who differ in generosity.

After the instructions for the DGP, participants had two practice rounds, once as the giver and once as the receiver. They then answered four comprehension check questions about the DGP (see Appendix B). Those who failed in this check (13 people, about 5% of participants) did not progress to the DGP phase of the study.

3.2.1.2.2 Cue manipulations in the DGP

Participants played ten rounds of the DGP, each with a different (sham) partner (they did not know how many rounds they would interact with others). The two IVs, *Group membership* and *Identifiability*, were manipulated in eight rounds. In these eight rounds, participants interacted with one of the eight sham partners shown in Table 2.1 in a random order. Namely, participants experienced both roles, giver and receiver, four rounds each, and each time they interacted with a sham partner with one of the four possible IV combinations: 2 (*ingroup* or *outgroup*) × 2 (*anonymous* or *identified*).

Table 2.1. Eight combinations of sham partners in study 2.

Group membership of partner	Participant's identifiability to partner	Partner's role in DGP
Ingroup	Anonymous	Giver
Ingroup	Anonymous	Receiver

Ingroup	Identified	Giver
Ingroup	Identified	Receiver
Outgroup	Anonymous	Giver
Outgroup	Anonymous	Receiver
Outgroup	Identified	Giver
Outgroup	Identified	Receiver

Participants were constantly reminded of the group membership of their partners as well as their identifiability during a DGP round. Throughout the round, a participant saw two silhouettes representing oneself and the partner. The silhouettes were painted in either red or blue based on the group(s) of the participant and the partner (see figure 2.1; see Appendix B for more details). The initials of the participant and the partner were shown beneath the silhouettes. When the participant was anonymous to the partner, their initials were replaced with "?.?."

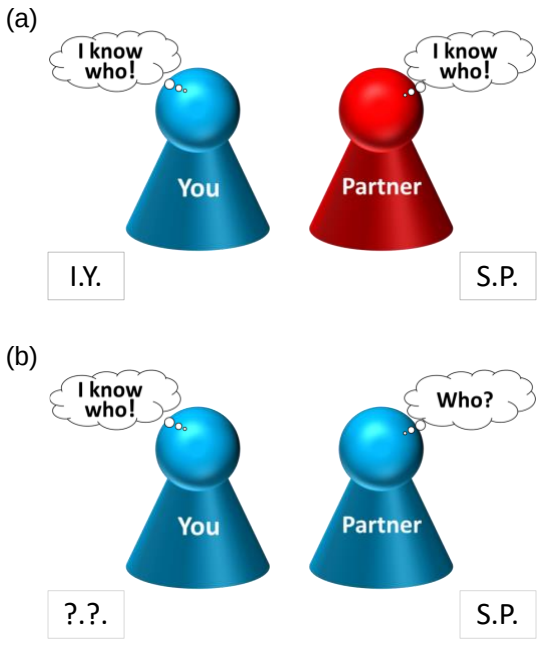


Figure 2.1. Examples of the two silhouettes representing the participant and the partner during a DGP round. (a) In the *identified* condition, both the participant and the partner saw each other's initials (in this example, the partner was an outgroup member). (b) In the *anonymous* condition, the partner did not see the participant's initials and saw "?.?." instead,

while the participant saw the partner's initials (in this example, the partner was an ingroup member).

When the participant was the receiver, the sham giver behaved stingily. The sham giver either (i) kept 100 points and gave the participant 50 points or (ii) kept 120 points and gave the participant 30 points. When the participant was the giver, the sham receiver conditionally punished the participant. If the participant gave the sham receiver 60 points or more, the sham receiver never punished the participant. If the participant gave the sham receiver 50 points or less, there was a 50% chance that the sham receiver paid 10 points to deduct 50 points from the participant.

In the remaining two rounds, there were no IV manipulations, and behaviors during these two rounds were not analyzed. The two rounds were set in between the eight rounds where DVs were measured (round 1 and 6). In these two rounds, participants interacted with anonymous sham givers: In contrast to the usual anonymous rounds, only the sham partner saw the participant's initials. The anonymous givers were always generous and gave either 70 or 80 points to the participant. One of the unknown generous partners was ingroup, the other was outgroup (random order). These two rounds were included to reduce participants' suspicion that their partners were not real people (in the eight rounds, partners were never anonymous, and they always shared stingily) by letting participants experience DGP rounds where (i) their partners became anonymous and (ii) sham givers do not behave stingily.

3.2.2 Study 2a Results

Data were analyzed using R 4.0.3 (R Core Team, 2020). I examined the effects of the two IVs, *Group membership* (ingroup vs. outgroup partners) and *Identifiability* (identified vs. anonymous), on the two DVs: *Giving* and *Punishment*.

3.2.2.1 What predicts giving?

When participants were the giver, they could share the 150-point endowment with their partner, the receiver. On average, participants gave the partner 58.3 points (*SD* = 29.9). The median was 70 points: 45% of givers gave the receiver about half of the endowment (16 % of them gave 80 points out of 150 points; 29% gave 70 points).

When it is likely that one is being recognized and assessed as a potential cooperation partner, investing in one's cooperative reputation would increase the probability of being chosen. I therefore predicted that motivations to cooperate will be up-regulated when it is likely that you are being evaluated as a partner, specifically, when (i) the current interaction partner is a member of your own group and (ii) you are identified as an individual, so that your reputations are trackable. I ran a linear mixed-effects model on *Giving* to test these predictions, using the lmer function in R package lme4 (Bates et al., 2015). Because the two IVs were within-subjects conditions and *Giving* responses were nested within participants, the mixed-effects model was employed to include a random intercept for each participant. The two IVs, (i) *Group membership* (ingroup receiver = 1, outgroup receiver = 0) and (ii) *Identifiability* (the participant was identified = 1, the participant was anonymous = 0), were entered as fixed effects. The interaction between the two IVs was also entered. A likelihood-

ratio test indicated that the model including fixed effects provided a better fit than a model without them ($\chi^2(3) = 21.6$, $p = 10^{-5}$).

There were no effects of *Identifiability* ($b = -1.87$, $p = .286$) or *Group membership* ($b = 2.53$, $p = .145$) on how many points participants gave to the partner, when controlling for the interaction between the two IVs. The interaction was significant: The effect of partner's group membership was dependent on whether the participant was identified or anonymous ($b = 5.19$, $p = .036$). Participants gave more points—on average, 5.19 points—to ingroup than to outgroup partners, but it was only when participants could be identified by their partners (figure 2.2). See Table 2.2 for the full model.

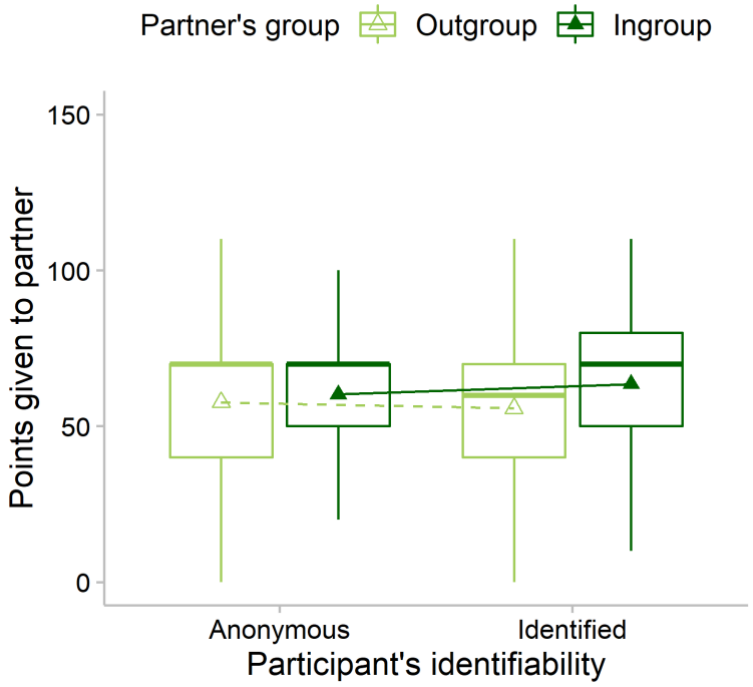


Figure 2.2. The effect of partner's group and participant's identifiability on giving in study 2a. Average points the participant gave to ingroup and outgroup partners when the participant was anonymous vs. identified by the partner. Boxplots show median and quartiles; triangles represent means.

Table 2.2. Fixed effects and intercept for giving in linear mixed effects model in study 2a.

Predictors	Estimate	*SE*	95% CI	*t*	*p*
Intercept	57.68	1.92	[53.91, 61.44]	30.11	10^{-16}
Identifiability (1 = identified, 0 = anonymous)	-1.87	1.75	[-1.56, 5.30]	-1.07	.286
Partner's group (1 = ingroup, 0 = outgroup)	2.53	1.75	[-0.9, 5.96]	1.45	.148
Interaction: Identifiability × Group	5.19	2.47	[0.33, 10.04]	2.10	.036

3.2.2.2 What predicts inflicting punishment?

When participants were the receiver, they could punish their partner's giving decision. Forty-three percent of the time, participants chose to inflict punishment on the sham partner, who gave the participant either 30 or 50 points out of 150.

Punishment may harm a reputation as a cooperator and lower the probability of being chosen as a partner. I therefore predicted that motivations to inflict punishment will be down-regulated in the presence of cues that you are being recognized and assessed as a potential partner, i.e., when (i) the current interaction partner is a member of your own group and (ii) you can be identified by the partner. To test these predictions, I ran a generalized linear mixed-effects model on the binary outcome, *Punishment* DV, using the mixed_model function in R package GLMMadaptive (Rizopoulos, 2022). This model also included a random intercept for each participant. The two IVs, *Group membership* and *Identifiability*, were entered as fixed effects. The interaction between the two IVs was insignificant and removed from the model. A likelihood-ratio test indicated that the model including the two IVs provided a better fit than a model without them (LRT (2) = 27.75, $p < 10^{-4}$).

Figure 2.3 shows the average probabilities of punishment. For ease of interpretation, I report the fixed effects as odds ratio (OR) (see table 2.3 for original estimate values). There was no effect of *Identifiability* of the participant (p = .189). The probability that participants

punished their stingy partners when they were anonymous was no different from the probability of punishing when participants could be identified (OR = 1.26, 95% CI = [0.89, 1.78]). However, there was a significant effect of *Group membership* of the partner ($p = 10^{-4}$). Participants were 2.47 times more likely to punish their partners when the partner was on a different team than when the partner was on the same team (OR = 2.47, 95% CI = [1.73, 3.53]).

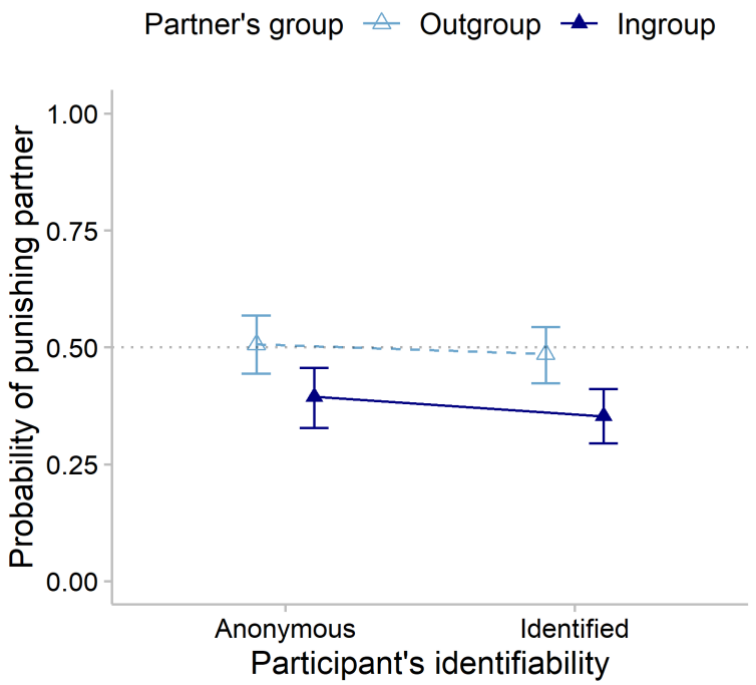


Figure 2.3. The effect of partner's group and participant's identifiability on punishment in study 2a. The probability of the participant punishing ingroup and outgroup partners when the participant was anonymous vs. identified by the partner. Triangles represent means; error bars show standard errors.

Table 2.3. Fixed effects and intercept for punishment in generalized linear mixed effects model in study 2a.

Predictors	Estimate	*SE*	95% CI	*z*	*p*
Intercept	0.08	0.22	[-0.35, 0.50]	0.36	.715
Identifiability (1 = identified, 0 = anonymous)	-0.23	0.18	[-0.58, 0.11]	-1.31	.189

Partner's group (1 = ingroup, 0 = outgroup)	-0.90	0.18	[-1.26, -0.55]	-4.95	10^{-4}

3.2.3 Study 2a Discussion

Group situations are hypothesized to evoke heuristics for reputation-based partner choice: The mind assumes that one is being recognized and evaluated as a potential cooperation partner when facing members of one's own group. Based on this hypothesis, study 2a tested (i) whether motivations to cooperate and punish are regulated by group membership of the partner and (ii) whether these effects can be attenuated when there is another cue that one is not being evaluated: being anonymous.

The results of study 2a support these predictions in motivations to cooperate. Participants' motivations to cooperate with their partners were up-regulated when they were interacting with ingroup partners rather than outgroup partners only when their partners could identify and evaluate them. In other words, when people knew that they were anonymous, such that their cooperative reputations were not at stake, they did not treat ingroup members more favorably than outgroup members. That is, ingroup-favoring cooperation disappeared when there was no need to manage your reputation so as to be chosen (or not to be excluded) by members of your own group. This is a conceptual replication of the findings by Yamagishi and colleagues (Jin & Yamagishi, 1997; Mifune et al., 2010; Yamagishi et al., 1999; Yamagishi & Mifune, 2008).

In contrast, motivations to inflict punishment on a stingy partner were affected only by group membership. Participants were more likely to punish outgroup partners than ingroup partners, regardless of whether they could be identified by their partners or not. This pattern

of ingroup-favoritism—perhaps better thought of as outgroup discriminating punishment—is consistent with many previous studies examining the effect of group membership on motivations to punish (e.g., Delton & Krasnow, 2017; Schiller et al., 2014).

There was no main effect of *Identifiability* on either giving or punishment behaviors; it influenced giving, but only for ingroup members. There are two possible explanations for why *Identifiability* did not affect punishment and affected giving only for ingroup members: (i) This could reflect how motivational systems detect and use cues for regulating motivations to cooperate versus punish, or (ii) the *Identifiability* manipulation—whether the partner saw the participant's initials—was too ambiguous as a cue of reputation-based partner choice. Participants in study 2a never experienced choosing a partner or being chosen as a partner. Furthermore, the participants were from anywhere—across 36 states in actuality—in the United States. There was a very slim chance that they would ever come across other participants in real life, let alone identify them. Indeed, this was what some participants reported that they felt during the interactions. After the study, I asked participants what they thought about the *Identifiability* manipulation ("When you were interacting with your partners, your partners sometimes knew who you were (they saw your initials). Sometimes they did not know who you were (they did not see your initials). Sometimes you did not know who they were (you did not see their initials). Did you have any thoughts about that?"). Several answered that they thought they were still *un*identifiable with their initials revealed (e.g., "I don't think the initials mattered because chances were I had no idea who the person was"; "Because initials are a relatively vague identifier of individuals, I didn't put much thought into it"). Studies 2b and 2c were conducted to address these limitations concerning the effect of the *Identifiability* cue.

3.3 Study 2b: College sample, online

In study 2b, two modifications were added to increase the saliency of the *Identifiability* cue. First, participants were told that, after interacting with several others, they would be allowed to choose who they would like to interact with in another DGP round (participants did not know how many rounds there would be). This gave participants a reason to believe that they were being assessed as a potential partner while playing DGP rounds. Participants were also encouraged to pay attention to their partners' initials—the only identifier that they could use to keep track of reputations of others. Along with these changes, participants were introduced to other players just before they started interacting with each other.

Second, study 2b was conducted in a smaller, more closely-knit community where initials are more likely to work as identifiers: undergraduate students studying on the same campus. The participants were recruited from the subject pool consisting of students who were taking either of two psychology courses. Because these courses were requirements to be admitted as a psychology major, most participants were in the same cohort taking other required classes together. Participants knew that other participants were fellow students from the same courses. In this sample, it was possible that participants personally knew each other and even could identify their partners by their initials (although, in reality, none of partners' initials were of real participants).

Regarding the first modification, participants were told that they could choose partners either only from ingroup members or only from outgroup members (between-subjects condition). The latter condition was created to test an additional prediction that stems from the hypothesis that group is a cue of reputation-based partner choice: When people are

explicitly told that they are going to choose partners only from outgroup members, motivations to cooperate with outgroup members will be up-regulated to manage their reputation in the eyes of the potential partners—outgroup members. That is, this additional manipulation tests whether experimentally changing the pool of potential partners will reduce ingroup favoritism.

3.3.1 Study 2b Methods

3.3.1.1 Participants

The study was conducted online. Participants were 223 English speakers in the United States (72% female, M_{age} = 19, SD_{age} = 1) recruited from an undergraduate psychology subject pool at University of California, Santa Barbara. Participants received a course credit for their participation. Five participants who did not provide consent to use their data were excluded from the analysis.

3.3.1.2 Design

The design was the same as study 2a with one exception: partner choice. After participants were introduced to their (sham) partners, they were instructed that later they would be able to choose whom they would like to interact with. Participants were also told that they might want to pay attention to their initials.

There were two conditions regarding from which group they could choose a partner. (i) In *Ingroup Partner Choice* condition, participants could choose partners only from the same team; (ii) in *Outgroup Partner Choice* condition, they could choose partners only from a different team. In both conditions, participants were explicitly instructed that the partner

choice was mutual—their potential partners would be also assessing them as partners. That is, they were instructed to believe that they were being assessed by their potential partners (by ingroup or outgroup members). They were told that they would be matched with their partner based on their preferences as well as the partner's preferences. (See Appendix B for details.)

After playing ten DGP rounds, participants were allowed to choose a partner. They were shown a list of (sham) partners with whom they had interacted in the previous ten rounds. In Ingroup Partner Choice condition, participants were only shown sham partners on the same team; in Outgroup Partner Choice condition, they saw only sham partners on a different team. Participants were instructed to rank the listed partners in order of preference.

Participants were automatically paired with their top choice. The chosen sham partner always played the giver and generously gave participants 80 points out of 150. After one round with the chosen partner, participants were told that there would be no more rounds.

3.3.2 Study 2b Results

The same analysis strategy was used as in study 2a. In addition to two IVs (*Group membership* and *Identifiability*) and their interaction, *Partner Choice* condition (Ingroup vs. Outgroup Partner Choice) was entered in the models.

3.3.2.1 Before interacting, some participants knew they would be choosing a partner from their own team whereas others knew they would be choosing a partner from the other team. Did that affect their giving or punishment behaviors?

No. Whether participants could choose partners only from their own group (n = 108) or a different group (n = 115) had no main effect on *Giving* (p = .691) or *Punishment* (p = .926). There were no interaction effects involving it either. The *Partner Choice* condition was therefore removed from the following models.

3.3.2.2 What predicts giving?

When participants were the giver, they could share the 150-point endowment with their partner. On average, participants gave the partner 56.89 points (SD = 30.15). The median was 70 points: 45% of givers gave the receiver about half of the endowment (out of 150 points, 12% gave 80 points; 32% gave 70 points). A likelihood-ratio test indicated that the model including fixed effects for the two IVs and interaction provided a better fit than a model without them ($\chi^2(3) = 21.6$, $p = 10^{-7}$).

There was no main effect of *Identifiability* on how many points participants gave to the partner (b = -1.84, p = .364). All else being equal, whether their initials were revealed to the partner did not affect how many points they shared with the partner. There was a main effect of *Group membership* (b = 4.13, p = .042). All else being equal, participants gave more points—4.13 points on average—to ingroup than to outgroup partners. However, this main effect was qualified by a significant interaction between the two IVs: The effect of partner's group membership was dependent on whether the participant was anonymous or identifiable to the partner (b = 6.68, p = .020). Participants gave more points—6.68 points—to ingroup

than to outgroup partners, but it was only when participants could be identified by their partners (figure 2.4). See Table 2.4 for the full model.

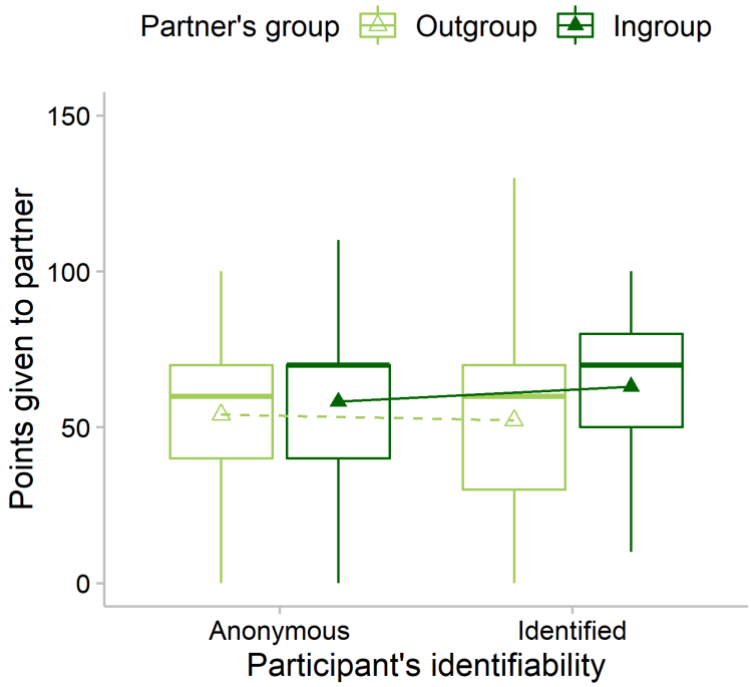


Figure 2.4. The effect of partner's group and participant's identifiability on giving in study 2b. Average points the participant gave to ingroup and outgroup partners when the participant was anonymous vs. identified by the partner. Boxplots show median and quartiles; triangles represent means.

Table 2.4. Fixed effects and intercept for giving in linear mixed effects model in study 2b.

Predictors	Estimate	*SE*	95% CI	*t*	*p*
Intercept	54.08	2.00	[50.16, 58.01]	27.06	10^{-16}
Identifiability (1 = identified, 0 = anonymous)	-1.84	2.03	[-5.81, 2.14]	-0.91	.364
Partner's group (1 = ingroup, 0 = outgroup)	4.13	2.03	[0.15, 8.10]	2.04	.042
Interaction: Identifiability × Group	6.68	2.86	[1.06, 12.30]	2.33	.020

3.3.2.3 What predicts inflicting punishment?

When participants were the receiver, they could punish their partner's giving decision. Forty-nine percent of the time, participants chose to inflict punishment on the sham partner, who gave the participant either 30 or 50 points out of 150.

The interaction between the two IVs was insignificant and removed from the model (p = .508). A likelihood-ratio test indicated that the model including the two IVs provided a better fit than a model without them (LRT (2) = 40.72, $p < 10^{-4}$).

There was a significant effect of *Identifiability* of the participant (p = .007): When participants were anonymous to the partner, they were 1.58 times more likely to inflict punishment on the stingy partner than when they could be identified (OR = 1.58, 95% CI = [1.13, 2.21]) (see table 2.5 for estimates and figure 2.5 for the average probabilities of punishment). There was a significant effect of *Group membership* of the partner as well ($p < 10^{-4}$): Participants were 2.69 times more likely to punish outgroup partners than ingroup partners (OR = 2.69, 95% CI = [1.90, 3.81]).

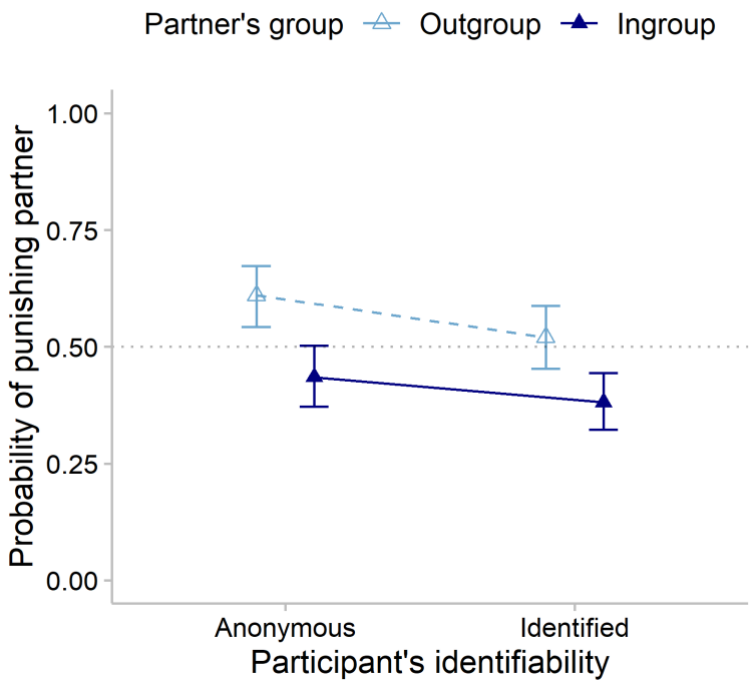


Figure 2.5. The effect of partner's group and participant's identifiability on punishment in study 2b. The probability of the participant punishing ingroup and outgroup partners when the participant was anonymous vs. identified by the partner. Triangles represent means; error bars show standard errors.

Table 2.5. Fixed effects and intercept for punishment in generalized linear mixed effects model in study 2b.

Predictors	Estimate	*SE*	95% CI	*z*	*p*
Intercept	0.63	0.20	[0.25, 1.02]	3.24	.001
Identifiability (1 = identified, 0 = anonymous)	-0.46	0.17	[-0.79, -0.12]	-2.68	.007
Partner's group (1 = ingroup, 0 = outgroup)	-0.99	0.18	[-1.34, -0.64]	-5.61	10^{-4}

3.3.3 Study 2b Discussion

Incentives to manage one's reputation as a cooperation partner were magnified in study 2b compared to study 2a. Participants in study 2b were allowed to choose partners, and the partner choice was mutual—they were instructed that they were being evaluated by their

partners during interactions. Therefore, when participants were not anonymous, they knew that their behaviors—whether they provide generously and whether they punish stinginess—could affect their reputations in their partners' eyes.

With these changes in mind, the effects of *Identifiability* and *Group membership* were reexamined in study 2b. *Identifiability* did significantly decrease motivations to inflict punishment on a stingy partner. Motivations to punish were down-regulated by a cue that one was being evaluated by potential partners. This was a separate effect from ingroup-favoring (outgroup discriminating) punishment: Motivations to inflict punishment were generally lower when the stingy partner was from one's own group than from a different group. This effect was not moderated by whether participants were anonymous or identifiable to the partner.

Study 2b also replicated the attenuation of ingroup-favoring cooperation that was found in study 2a. Ingroup-favoring cooperation was reduced when participants were anonymous—another replication of Yamagishi et al. (Jin & Yamagishi, 1997; Mifune et al., 2010; Yamagishi et al., 1999; Yamagishi & Mifune, 2008). In contrast to study 2a, the main effect of *Group membership* was significant in study 2b: Motivations to cooperate were generally higher when interacting with ingroup partners than when with outgroup partners, controlling for the interaction.

In addition to engaging in partner choice, study 2b participants were students taking the same courses and studying on the same campus. The concern for managing one's reputations could be stronger among them than among participants of study 2a, who were recruited from a large crowd-sourcing website. Being anonymous might not have been

enough to erase the reputation concern when interacting with ingroup members, who could be classmates and neighbors in real life.

Additionally, before interacting with partners, some participants in study 2b knew that they would be choosing a partner from their own group, while others knew that they would be choosing a partner from an outgroup. This manipulation did not affect either cooperation or punishment behaviors. Experimentally changing the pool of potential partners from ingroup to outgroup did not reduce ingroup favoritism.

Study 2c was conducted to further examine the robustness of these effects by further emphasizing the possibility of being evaluated as a partner.

3.4 Study 2c: College sample, in-person

Study 2c was conducted in-person to further highlight the possibility of being evaluated as a partner. Participants could see other participants taking part in the same study session, believing that they were interacting with one another. In this setting, initials were not the only identifier. Participants knew that their partners saw their face—an ancestrally-reliable cue to identify—and could keep track of other individuals. Considering that participants were classmates in real life, participants in study 2c may have felt as if they could be actually identifiable with their faces revealed to their partners (although their face were never paired with their initials during the study).

3.4.1 Study 2c Methods

Study 2c was almost identical to study 2b except that study 2c was conducted in-person. Up to five participants took part in the study at the same time in a university lab, and it was

stressed that they were interacting with others in the room. Before the session began, a research assistant announced that they might have to wait while other participants make their decisions and that they would not be allowed to leave until all participants finished the study.

Participants used desktop computers to take part in the study. During the session, participants could see each other, but they could not see the computer displays of other participants because of dividers between computers.

3.4.1.1 Participants

The study was conducted in a psychology lab at University of California, Santa Barbara. Participants were 259 English speakers in the United States (72 % female, M_{age} = 19, SD_{age} = 1) recruited from an undergraduate psychology subject pool. Participants received a course credit for their participation. Three participants who did not provide consent to use their data were excluded from the analysis.

3.4.1.2 Design

The design was exactly same as study 2b with one exception: a *Universal Partner Choice* condition was included. In this condition, participants could choose a partner from either group. After ten DGP rounds, participants in Universal Partner Choice condition were given a list of (sham) partners on the same team and a different team and allowed to choose partners regardless of group membership. Participants were randomly assigned to one of the three partner choice conditions: Universal, (only) Ingroup, or (only) Outgroup Partner Choice. The Universal Partner Choice condition was added to further examine whether the

Outgroup Partner Choice condition will reduce ingroup favoritism by heightening reputation concern *only* toward outgroup partners, compared to other conditions where there is reputation concern when interacting with ingroup partners.

3.4.2 Study 2c Results

The same analysis strategy was used as in study 2b.

3.4.2.1 Before interacting, participants knew that they would be choosing a partner from the same team, from a different team, or from either team. Did that affect their giving or punishment behaviors?

No. Whether participants could choose partners only from their own group (n = 114), only from a different group (n = 112), or from either group (n = 33) had no effects on the two DVs, *Giving* or *Punishment*. The *Partner Choice* condition was removed from the models below.

3.4.2.2 What predicts giving?

When participants were the giver, they could share the 150-point endowment with their partner. On average, participants gave the partner 59.82 points (SD = 28.24). The median was 70 points: 46% of givers gave the receiver close to half of the endowment (out of 150 points, 17% gave 80 points; 29% gave 70 points). A likelihood-ratio test indicated that the model including fixed effects for the two IVs and interaction provided a better fit than a model without them ($\chi^2(3) = 54.45$, $p = 10^{-4}$).

There was no main effect of *Identifiability* on how many points participants gave to the partner (b = 0.50, p = .775). All else being equal, whether their initials were revealed to the

partner did not affect how generously they behaved. There was a main effect of *Group membership* (b = 3.63, p = .039). All else being equal, participants gave more points—3.63 points on average—to ingroup than to outgroup partners. However, this main effect was qualified by an interaction between the two IVs: The effect of partner's group membership was dependent on whether the participant was anonymous or identifiable to the partner (b = 7.53, p = .003). Participants gave more points—on average, 7.53 points—to ingroup than to outgroup partners, but it was only when participants could be identified (see figure 2.6). See Table 2.6 for the full model.

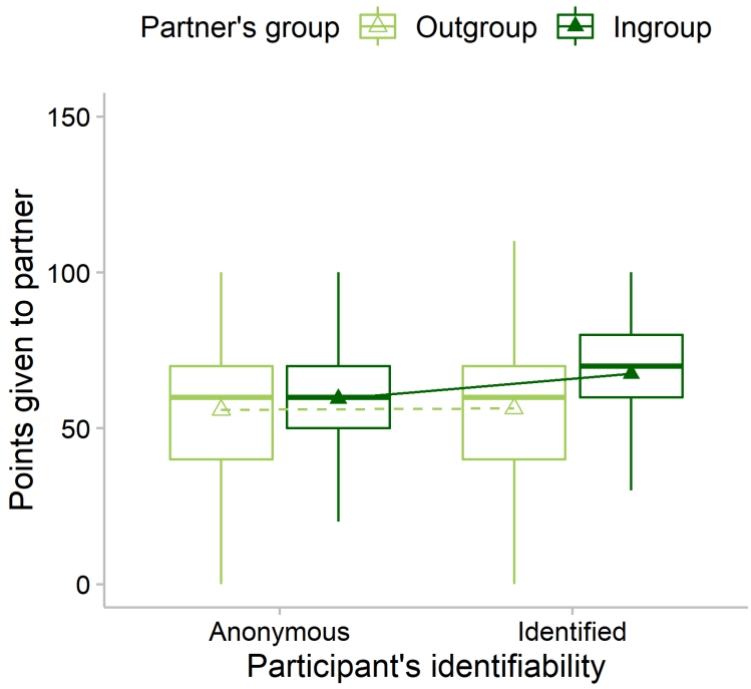


Figure 2.6. The effect of partner's group and participant's identifiability on giving in study 2c. Average points the participant gave to ingroup and outgroup partners when the participant was anonymous vs. identified by the partner. Boxplots show median and quartiles; triangles represent means.

Table 2.6. Fixed effects and intercept for giving in linear mixed effects model in study 2c.

Predictors	Estimate	*SE*	95% CI	*t*	*p*
Intercept	55.87	1.73	[52.47, 59.26]	32.30	10^{-16}
Identifiability (1 = identified, 0 = anonymous)	0.50	1.76	[-2.95, 3.95]	0.29	.775
Partner's group (1 = ingroup, 0 = outgroup)	3.63	1.76	[0.18, 7.08]	2.07	.039
Interaction: Identifiability × Group	7.53	2.48	[2.65, 12.40]	3.03	.003

3.4.2.3 What predicts inflicting punishment?

When participants were the receiver, they could punish their partner's giving decision. Forty-eight percent of the time, participants chose to punish the stingy (sham) partner.

The interaction between the two IVs was insignificant and removed from the model (p = .496). A likelihood-ratio test indicated that the model including the two IVs provided a better fit than a model without them (LRT (2) = 43.71, $p < 10^{-4}$).

There was a significant effect of *Identifiability* of the participant (p = .004): When participants were anonymous, they were 1.58 times more likely to inflict punishment on the stingy partner than when they could be identified (OR = 1.58, 95% CI = [1.16, 2.17]) (see table 2.7 for estimates and figure 2.7 for the average probabilities of punishment). There was a significant effect of *Group membership* of the partner as well ($p < 10^{-4}$): Participants were 2.58 times more likely to punish outgroup partners than ingroup partners (OR = 2.58, 95% CI = [1.87, 3.56]).

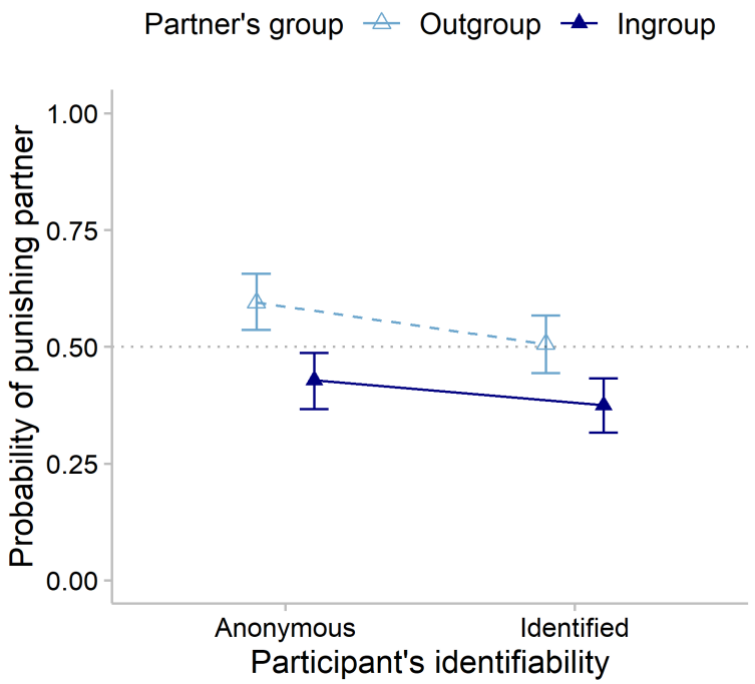


Figure 2.7. The effect of partner's group and participant's identifiability on punishment in study 2c. The probability of the participant punishing ingroup and outgroup partners when the participant was anonymous vs. identified by the partner. Triangles represent means; error bars show standard errors.

Table 2.7. Fixed effects and intercept for punishment in generalized linear mixed effects model in study 2c.

Predictors	Estimate	*SE*	95% CI	*z*	*p*
Intercept	0.54	0.18	[0.18, 0.90]	2.97	.003
Identifiability (1 = identified, 0 = anonymous)	-0.46	0.16	[-0.77, -0.15]	-2.89	.004
Partner's group (1 = ingroup, 0 = outgroup)	-0.95	0.16	[-1.27, -0.63]	-5.78	10^{-4}

3.4.3 Study 2c Discussion

Study 2c replicated study 2b. First, motivations to cooperate were up-regulated only during an interaction with someone from your own group who could identify you. Ingroup-favoring cooperation was attenuated when participants believed that they were anonymous,

again conceptually replicating the findings by Yamagishi and colleagues (e.g., Yamagishi et al., 1999). Second, motivations to inflict punishment on a stingy partner were separately down-regulated by two cues indicating the possibility of being evaluated as a cooperation partner. The probability of punishing a stingy partner was lowered either when (i) the partner was an ingroup member or when (ii) the partner could keep track of your reputations. Additionally, experimentally changing the pool of potential partners—choosing partners from ingroup vs. outgroup vs. either—did not reduce ingroup favoritism in terms of cooperation or punishment, replicating study 2b. Alternatively, by only letting participants experience choosing a partner from the outgroup once, the current manipulation may have failed to change their assumption that ingroup members are the pool of potential partners.

Reputation management could be a more realistic concern for participants in study 2c than in study 2b. With their faces and initials known (although they were not paired), it was not impossible that their partners, classmates studying on the same campus, could identify who they were. Yet, the heightened reputation concern did not change the results. Potentially, the saliency of *Identifiability* was high enough to indicate the possibility of being evaluated in study 2b, which differed from study 2c only in that it was run online. Even interacting online without physically seeing each other, other participants—classmates or neighbors—were potential partners in real life. Plus, even though the human face is a critical cue to keep track of social relationships in reality, it was irrelevant to partner choice decisions in the experiment, where initials were the only identifier. For these reasons, seeing potential partners in-person might not have increased the estimated possibility of being assessed as a partner.

3.5 Study 2 Pooled-analysis

Three studies found significant effects of two cues of reputation-based partner choice, *Identifiability* and *Group membership*, on motivations to cooperate and punish. To check the robustness of these results and to estimate the overall effect sizes, I conducted a pooled (meta) analysis of 3 studies.

3.5.1 Pooled-analysis Methods

Because the three studies differ in populations (Prolific vs. college) and data collection methods (online vs. in-person), I first estimated the overall effect sizes using random effect models and then considered the variation in each effect size distribution by using indicators of heterogeneity. I used the restricted maximum likelihood estimator (Viechtbauer, 2005) for the *Giving* DV, using the metacont function in R package meta (Balduzzi et al., 2019). To estimate the odds ratio for *Punishment*, I used the Paule-Mandel estimator (Paule & Mandel, 1982) with metabin function in meta (Balduzzi et al., 2019). Knapp-Hartung adjustments (Knapp & Hartung, 2003) were used to calculate the confidence intervals around the overall effects.

3.5.2 Pooled-analysis Results

3.5.2.1 What predicted giving across 3 studies?

I first estimated the overall effect sizes of *Identifiability* and *Group membership* on how many points participants gave to their partners (Giving DV). A random effect model suggests that, across 3 studies, there was no significant main effect of *Identifiability*

(identifiable vs. anonymous) on how generous participants were (Hedge's g = 0.08, 95% CI = [-0.09, 0.25]). But *Group membership* (ingroup vs. outgroup) had a small positive main effect: Participants gave more points to ingroup than outgroup partners (g = 0.23, 95% CI = [0.10, 0.35]). There was no heterogeneity in the effect size distributions (T^2 = 0.0007, 95% CI = [0.0000, 0.18] for *Identifiability*; 0.00, 95% CI = [0.0000, 0.09] for *Group membership*) and these variations cannot be explained by the difference between studies (I^2 = 11.6%, 95% CI = [0.0%, 90.8%] for *Identifiability*; 0.00%, 95% CI = [0.0%, 89.6%] for *Group membership*).

All three studies found a significant interaction effect between *Identifiability* and *Group membership* on the *Giving* DV. To get a better estimate for the interaction, I pooled data from the three studies and ran a linear mixed-effects model. Controlling for which study the data came from, there was a significant interaction between *Identifiability* and *Group membership* (b = 6.49, $p = 10^{-4}$). Across 3 studies, participants gave more points—on average, 6.49 points—to ingroup than to outgroup partners when participants could be identified by the partners. As suggested by there being no evidence of heterogeneity, how generous participants behaved did not differ across these studies (p = .186 for the effect of data coming from study 2a compared to study 2b as a baseline; .125 for the effect of data coming from study 2c compared to 2b). See Table 2.8 for the full model.

Table 2.8. Fixed effects and intercept for giving in pooled data from three studies (2a, 2b, and 2c).

Predictors	Estimate	*SE*	95% CI	*t*	*p*
Intercept	54.07	1.59	[50.95, 57.19]	33.96	10^{-16}
Data from study 2a: Prolific (no partner choice)	2.54	1.92	[-1.23, 6.30]	1.32	.186
Data from study 2c: College (in-person)	2.92	1.90	[-0.81, 6.65]	1.54	.125
Identifiability (1 = identified, 0 = anonymous)	-1.01	1.26	[-3.48, 1.46]	-0.80	.424
Partner's group (1 = ingroup, 0 = outgroup)	3.42	1.26	[0.94, 5.89]	2.71	.007
Interaction: Identifiability × Group	6.49	1.79	[2.99, 9.99]	3.64	10^{-4}

3.5.2.2 What predicted inflicting punishment across 3 studies?

The overall effect sizes of *Identifiability* and *Group membership* on the probability of inflicting punishment (*Punishment* DV) were estimated. Across 3 studies, participants were (i) 1.27 times more likely punish stingy partners when the participants were anonymous than identifiable (OR = 1.27, 95% CI = [1.02, 1.58]) and (ii) 1.78 times more likely to punish outgroup than ingroup partners (OR = 1.78, 95% CI = [1.50, 2.11]). There was no heterogeneity in the effect size distributions (T^2 = 0.00, 95% CI = [0.0000, 0.29] for *Identifiability*; 0.00, 95% CI = [0.0000, 0.17] for *Group membership*); these variations cannot be explained by the difference between studies (I^2 = 0.00%, 95% CI = [0.0%, 89.6%] for both *Identifiability* and *Group membership*).

3.5.3 Pooled-analysis Discussion

I aggregated data from 723 participants across 3 experiments to examine whether motivations to cooperate and punish were robustly affected by two cues of reputation-based partner choice, group membership and identifiability, each suggesting that one may be being recognized and evaluated as a potential cooperation partner. First, motivations to cooperate were up-regulated by group membership *conditionally* with identifiability. Across 3 experiments, people behaved more generously toward ingroup members than outgroup partners, but this effect was attenuated when people believed that they were anonymous and could not be evaluated as a partner. Second, motivations to inflict punishment were *independently* down-regulated by group membership and identifiability: People were less

likely to inflict punishment on a stingy partner either (i) when the partner is from their own group or (ii) when the partner could identify them and assess their reputations. In short, the pooled analyses revealed that the two situational cues of reputation-based partner choice have robust effects on motivations to cooperate and punish.

3.6 Study 2 General discussion

Three experiments examined how motivations to cooperate and punish are regulated in response to two cues suggesting that one is being recognized and evaluated as a potential cooperation partner. Across three experiments, those cues up-regulated motivations to cooperate and down-regulated motivations to punish, but there were differences in how the two cues interacted with each other in affecting these two different motivations.

3.6.1 Ingroup-favoring cooperation as a byproduct of reputation management

I employed the two cues to test the hypothesis that the mind considers members of your own group as a pool of potential cooperation partners, such that group situations elicit concerns to acquire a reputation as a cooperator. This hypothesis predicts that (i) motivations to cooperate will be up-regulated toward a member of your own group, who is likely to consider you as a potential partner and to assess your reputations. But (ii) when the potential partner does not know your identity and thus cannot form or keep track of your reputations, the motivational up-regulation will be suppressed.

In contrast, theories of social identity and parochialism produce a contrasting prediction about (ii). According to these theories, favoring ingroup is to positively differentiate your

own group from other groups so that you can enhance your social identity as a group member—it is not for investing in your individual reputation as a cooperator. This alternative hypothesis predicts that (i) people will be motivated to provide more benefits to ingroup members than to outgroup members, but it will be (ii) indiscriminate, regardless of whether they are anonymous or identified.

As predicted by the hypothesis that group cues a pool of potential partners, (i) people generally cooperated more with ingroup members than outgroup members, (ii) but this ingroup-favoring cooperation was either eliminated or attenuated when anonymity assured that there was no need to manage your reputation. Merely sharing group membership did not trigger ingroup-favoring cooperation but reputation concern with ingroup members did, a conceptual replication of the findings by Yamagishi et al. (Jin & Yamagishi, 1997; Mifune et al., 2010; Yamagishi et al., 1999; Yamagishi & Mifune, 2008).

The current study further extends their line of research. In the previous studies by Yamagishi and colleagues, reputational concern was manipulated via knowledge about group membership: In the control condition, an ingroup partner knew that a participant was from the same group; in the experimental, "no reputation concern" condition, the ingroup partner did not know that. Ingroup-favoring cooperation was observed mostly in the control condition, where it was plausible that you were being assessed as a cooperation partner. There were no cues to the identity of individuals in either condition.

By contrast, group membership was known by both partners in the three experiments reported here. In the anonymous condition here, an ingroup partner knew that the participant was an ingroup member, despite not knowing the participant's individual identity (i.e., initials)—this is the same situation as the control condition in the Yamagishi studies, where

no one knew anyone's name regardless of conditions. Unlike the Yamagishi studies, however, when participants were anonymous, their motivations to cooperate with their ingroup partners were not particularly high. Instead, motivations to cooperate with the ingroup partner were up-regulated when the ingroup partner knew the participant's identity and therefore could keep track of the participant's reputations (the identified condition). This finding adds to the literature by newly demonstrating that, to be motivated to treat ingroup members favorably, one needs not only to be recognized as *an ingroup member* but also to be identifiable and trackable as *an individual*.

More importantly, these results speak against the hypothesis that ingroup favoritism is indiscriminate, as implied by theories of social identity and parochialism. If motivational mechanisms were designed to positively differentiate one's own group from other groups regardless of the potential for cooperating with ingroup members, then being anonymous or identifiable should have no effect on motivations to deliver benefits to them. Instead, the results herein provide a clear support for the hypothesis that motivational systems are designed for managing one's reputation as a cooperator. The current results indicate that ingroup-favoring cooperation is not for the sake of the group but a byproduct of reputation management.

3.6.2 Is out-group discriminating punishment a product of reputation management?

The current experiments also examined whether these two cues affect motivations to inflict punishment. It was predicted that the possibility of being recognized and evaluated as a cooperation partner would down-regulate motivations to punish, because inflicting

punishment might harm one's cooperative reputation and lower the probability of being chosen as a partner. As predicted, people were less likely to punish a stingy partner when the partner was a member of one's own group than when the partner was from a different group. The current result suggests that group serves as a cue of reputation-based partner choice to systems regulating motivations to punish and elicits reputation concern to be chosen as a partner.

Several studies similarly show that people punish outgroup members more harshly than ingroup members. But the current experiments shed new light on the potential function of punishment regarding reputation management in group situations. In many past studies examining punishment in group contexts, the punisher had no incentives to manage their cooperative reputation in the eyes of an outgroup cheater (Bernhard, Fehr, et al., 2006; Bernhard, Fischbacher, et al., 2006; Delton & Krasnow, 2017; Goette et al., 2012; Guo et al., 2020; Jordan et al., 2014; Schiller et al., 2014; Yudkin et al., 2016). In contrast, about half of the participants in studies 2b and 2c had incentives to acquire a cooperative reputation among their outgroup members. This is because participants in the *Outgroup Partner Choice* condition could only choose partners from outgroup members, and at the same time, they were instructed to believe that they were being assessed by their outgroup partners. However, this condition did not affect the probability of punishing outgroup partners, indicating that the lack of concern for harming one's cooperative reputation may not explain why people engage in outgroup-discriminating punishment.

Nonetheless, motivations to punish were down-regulated when people believed that they could be identified compared to when they were anonymous. This effect was independent of the effect of outgroup-discriminating punishment discussed above. The current results

suggest that punishers do experience reputation concern when others can identify them and keep track of their reputations, and that down-regulates motivations to punish. However, the concern might not be about hurting one's reputation as a cooperator. For example, participants may have been afraid of retaliation when they were identifiable. It is not common that people in these populations (online crowd workers and college students) retaliate after they are punished (Arai et al., 2022; Bone et al., 2015, 2016).

The current data cannot rule out the possibility that the patterns of punishment were produced by processes of social identity or parochialism. These explanations posit that people have a general bias against outgroup members (Bernhard, Fischbacher, et al., 2006; Delton & Krasnow, 2017). It was possible that participants were more likely to punish outgroup than ingroup partners as a part of their general tendency to treat outgroup members poorly, irrespective of reputation management. Yet, this kind of explanation alone cannot provide a clear reason why being anonymous down-regulated punitive motivations toward outgroup and ingroup partners equally, rather than punishing outgroup members regardless of anonymity, or suppressing only outgroup punishment by assuming that outgroup members were more vengeful than ingroup members.

Future research could separate the effect of reputation management from that of social identity or parochialism, and further investigate whether outgroup-discriminating punishment is caused by lack of reputation concern. Both may be done by improving the experimental design regarding partner choice and letting participants actually compete for being chosen by desirable partners. The current experiments did not provide participants the experience of reputation-based partner choice. Partner choice conditions were not part of study 2a; because the opportunity to exercise partner choice came at the end of studies 2b

and 2c, participants did not experience choosing or being chosen before they had decided whether to punish (outgroup) partners. Similarly, they did not discover whether their behaviors or reputations affected whether they were chosen by desirable partners. These factors might explain why the *Outgroup Partner Choice* condition, where the outgroup was defined as the functional "ingroup"—the pool of partners—did not affect outgroup-discriminating punishment. A better manipulation in the future may be providing participants with experiences of being (not) chosen by outgroup partners—ideally multiple times—before measuring their motivations to punish ingroup vs. outgroup partners.

3.6.3 The targets of cooperative and punitive reputations

The ways two cues of reputation-based partner choice (*Group membership* and *Identifiability*) interacted might indicate the functions of acquiring cooperative versus punitive reputations in group contexts. First, there was a significant interaction effect between the two cues on motivations to cooperate: Motivations to cooperate were up-regulated only toward an *ingroup* member (vs. *outgroup* member) when participants were *identifiable* (vs. *anonymous*). Indeed, if group marks a pool of potential partners, it would be most advantageous to up-regulate motivations to cooperate toward a potential partner (*ingroup* member) who can keep track of your reputations (you are *identifiable*). But if observers are unlikely to consider you as a potential partner (*outgroup* members), being recognized as cooperative would not result in forming a cooperative relationship. The interaction effect suggests that up-regulating motivations to cooperate and investing in a cooperative reputation may have been beneficial mostly within the boundary of one's own group. That is, in the ancestral environment where motivational systems evolved, a

reputation as a cooperator might have been like a currency only used in one's own group, which became useless when interacting with people outside.

In contrast, the two cues had no interaction on regulating motivations to punish: Regardless of whether a partner was from the same group or a different group, motivations to punish were regulated in the same way in response to identifiability. This might indicate that up-regulating motivations to punish has been equally protective inside and outside one's group. A punitive reputation might have worked like a universal currency, discouraging cheaters in your own group from approaching while scaring off ill-intentioned outsiders looking for prey. Perhaps the current result—people were more likely to punish outgroup than ingroup partners—indicates that the latter function was more important than the former for protecting yourself as well as your fellow cooperation partners from those who are unlikely to form cooperative relationships with you.

3.6.4 Conclusion

Three experiments demonstrated that the possibility of being recognized and evaluated as a potential partner regulates motivations to cooperate and punish in dyadic social exchange. Motivations to cooperate were up-regulated when acquiring a reputation as a valuable cooperator would be most advantageous—during interactions with members of your own group, who are likely to consider you as a potential cooperation partner, but only when they could keep track of your reputations. Motivations to punish a stingy partner were down-regulated when acquiring a reputation for being punitive could be harmful—either when interacting with those who would regard you as a potential partner (ingroup members) or when your reputations were trackable. Overall, the results suggest that motivational

systems are designed for managing reputations to attract desirable partners in your own group while deterring undesirable partners regardless of group boundaries.

Chapter 4: Study 3. The trade-off between cooperative and punitive reputations [5]

Results of studies 1 and 2 were consistent. The pattern was that motivations to punish are down-regulated when it is estimated that (i) others have many opportunities to find new partners (study 1) or (ii) they are evaluating you as a cooperation partner (study 2). That is, motivations to punish were down-regulated when you were likely to be in competition to be chosen as a partner. A functional analysis of the motivational calibration suggests that punishment may harm one's reputation as a cooperation partner.

However, the reputational consequences of punishment are still unknown. Does punishing a cheating or stingy partner actually harm one's reputation as a cooperator? Does it lower the probability of being chosen by desirable partners? A result from study 1 provides a partial answer: Punishment drove away partners—cheaters and cooperators equally. Participants who were punished were more likely to switch partners than those who were not, and this effect was pronounced in participants who did reciprocate. But study 1 cannot address why those participants left the punitive partner. One possibility is that they wanted to run away from someone who harmed them personally. Another possibility is that they did not want to interact with people who punish others—regardless of whether they were the target of punishment. That is, it is unclear whether people avoid a punisher without

[5] This work has been submitted for publication as: Arai, S., Tooby, J., & Cosmides, L. (under review). Why punish cheaters? Those who withdraw cooperation enjoy better reputations than punishers, but both are viewed as difficult to exploit. Evolution and Human Behavior.

having had a personal experience of being punished by that person. Study 3 investigates whether inflicting punishment damages one's reputation as a dyadic cooperation partner in the eyes of a third-party observer. It compares three reactions to a cheater: punishment, withdrawing cooperation, and not sanctioning at all.

4.1 Study 3 Introduction

Negatively sanctioning cheaters promotes cooperation. But there are two ways of sanctioning partners who fail to reciprocate: by withdrawing cooperation or inflicting punishment. Punishment—inflicting a cost that reduces the payoff of a cheater—has been shown to successfully sustain cooperation (Fehr & Gächter, 2000, 2002; Yamagishi, 1986). But inflicting punishment is sometimes costly to the punisher as well (Clutton-Brock & Parker, 1995), leading theorists to ask how selection could have favored punishment as a means of sanctioning cheaters (Panchanathan & Boyd, 2004; Tooby et al., 2006).

Several researchers have proposed that the cost of inflicting punishment can be recouped if punishers acquire reputations as better cooperative partners than non-punishers, thereby attracting (or retaining) more rewarding partners for future interactions (Barclay, 2006; Horita, 2010; Kiyonari & Barclay, 2008; Ozono & Watabe, 2012; Raihani & Bshary, 2015a). Tests of this hypothesis have generated mixed results. Some studies found that punishers were seen as more trustworthy and received more benefits than non-punishers (Barclay, 2006; dos Santos et al., 2013; Jordan et al., 2016; Nelissen, 2008; Raihani & Bshary, 2015b), but others found that punishers were seen as less trustworthy and reaped no advantage over non-punishers (Balafoutas et al., 2014; Barclay & Raihani, 2016; Bone et al.,

2016; Fehr & Rockenbach, 2003; Kiyonari & Barclay, 2008; Przepiorka & Liebe, 2016). Many variables could account for these conflicting results (Horita, 2010; Mifune et al., 2020; Ozono & Watabe, 2012; Raihani & Bshary, 2015a). The reputational consequences of punishment may vary with context, for example: Punishment is the only method of selectively sanctioning cheaters in group cooperation, but not in dyadic cooperative exchanges (Tooby et al., 2006). For this reason, our research focused on negative sanctions in one context: dyadic cooperation.

Study 3 investigated the reputational consequences of three possible responses a cooperator could have to a partner's failure to reciprocate: inflicting punishment, withdrawing cooperation, and not sanctioning at all. Conditional cooperation is evolutionarily stable against strategies that defect (Tooby et al., 2006; Trivers, 1971; Williams, 1966), but many negative sanctions can incentivize a defecting partner to cooperate. Punishment does so by reducing the immediate payoff the partner gains by defecting. An alternative sanctioning strategy is to withdraw the benefits of cooperation: One can refrain from delivering additional benefits until the partner resumes cooperation (as TIT FOR TAT does; Axelrod & Hamilton, 1981) or switch to a more rewarding partner until the defector reforms (Hammerstein & Noë, 2016; Tooby et al., 2006).

Very few studies have directly compared behavior in response to these two negative sanctions: punishing versus withdrawing cooperation (for an exception, see Barclay & Raihani, 2016). Moreover, we can find no studies of the reputational consequences of withdrawing cooperation, even though this was the most widely studied method of sanctioning in the early literature on the evolution of cooperation (e.g., Axelrod, 1984). The

reputation attributed to those who withdraw cooperation has not been compared to that of punishers—or to the reputation of those who do not sanction at all.

We examined how these two methods of sanctioning influence the inferences observers make about the sanctioner's character and traits—the various reputations (plural) that observers attribute to the sanctioner. The colloquial use of *reputation* implies a unitary dimension: Your reputation can become better or worse. But people are routinely evaluated on many different traits: Alex may have a reputation for being generous, a reputation for being lazy, and a reputation for being vengeful, for example. These need not merge to form a single "reputation." And, even if they do, these separate reputations should remain stored in the observer's memory, because which is most relevant depends on the situation a decision-maker is facing (Klein et al., 2002). Indeed, research on social cognition shows that the mind spontaneously infers many different traits rapidly, even from thin information (Funder & Sneed, 1993; Klein et al., 2009), and stores summary representations of each (Klein et al., 2009).

Here we test two previously unexamined hypotheses about the inferences people draw from a cooperator's response to a partner who defects. The first hypothesis regards the reputations of cooperators who respond by imposing negative sanctions: withdrawers and punishers. In the two studies reported herein, *withdrawer* refers to a cooperator who sanctions by not providing benefits to the defector in the next round, and *punisher* refers to a cooperator who sanctions by removing resources from the defector in the next round. We propose that withdrawers will acquire reputations for being more cooperative than punishers—they will be seen as, e.g., more generous, trustworthy, and forgiving. As a result,

observers will prefer withdrawers to punishers as potential partners (Barclay, 2013; Roberts et al., 2021).

Why? Both withdrawers and punishers signal a willingness to sanction a defection, but withdrawers do so without reducing the payoff to a potentially well-intentioned cooperator. Even reliable cooperative partners will sometimes fail to reciprocate due to mistakes or bad luck (Delton et al., 2012); deciding whether a failure reveals a disposition to cheat or a mistake is a judgment made under uncertainty. Because they are robust to mistakes, strategies that require more evidence before sanctioning a partner, such as TIT FOR TWO TATS, outcompete strategies that sanction immediately in agent-based simulations (Axelrod, 1984). As a result, they maintain cooperation with a partner instead of triggering cycles of mutual defection.

In study 3a, sanctions are immediate in both cases and neither cooperator donates resources to their partner in the round following defection. But punishers take back what they gave whereas withdrawers do not. The partner—who may have made a mistake—retains the payoff provided by the withdrawer in the first round. This should lead observers to see the withdrawer as more generous and less vengeful than the punisher.

The second hypothesis addresses the reputational cost of *not* imposing negative sanctions when a partner defects. In both studies, *non-sanctioners* are cooperators who respond to defection by continuing to provide benefits to their partner. We propose that non-sanctioners will acquire a reputation for being more exploitable than those who impose negative sanctions, whether the sanctioners are punishers or withdrawers.

Why? Motivations to sanction defections could have been favored by selection if their average effect was to either increase benefits to the sanctioner and/or prevent losses.

Previous research on the reputational consequences of sanctioning has focused on whether punishers gain more benefits from cooperation than non-sanctioners do (Balafoutas et al., 2014; Barclay, 2006; Barclay & Raihani, 2016; Bone et al., 2016; dos Santos et al., 2013; Fehr & Rockenbach, 2003; Horita, 2010; Jordan et al., 2016; Kiyonari & Barclay, 2008; Mifune et al., 2020; Nelissen, 2008; Ozono & Watabe, 2012; Przepiorka & Liebe, 2016; Raihani & Bshary, 2015a, 2015b). But only a handful of studies have examined the possibility that sanctioning protects the sanctioner from further losses (Delton & Krasnow, 2017; Hilbe & Traulsen, 2012; Krasnow et al., 2016; Yamagishi et al., 2009). The few studies that do suggest that motivations to sanction were designed to deter further maltreatment by the defector or other observers. In this view, the cost of *not* sanctioning defections is gaining a reputation for being exploitable, which invites mistreatment. If selection for preventing losses designed motivations to sanction defectors, then observers will view non-sanctioners as more exploitable than sanctioners.

We tested these two hypotheses by having participants observe how a cooperator responded to a failure to reciprocate. After, they made inferences about the character and traits of withdrawers, punishers, and non-sanctioners.

- H1: Withdrawers will be evaluated more favorably as a cooperation partner than punishers.
- H2: Sanctioners—withdrawers and punishers—will be evaluated as less exploitable than non-sanctioners.

Inflicting punishment was cost-free in study 3a and costly in study3b.

4.2 Study 3a

In most theoretical and empirical work on the reputational consequences of punishment, the punisher pays a cost to reduce the payoff of a defector.[6] But reducing that payoff need not be costly if what is taken from the defector goes to the punisher. The punisher can recoup the investment lost by the defector's failure to reciprocate (or take more, to impose an additional penalty for cheating). In study 3a, there is no cost to sanctioning, but punishers reclaim what they lost and withdrawers do not. It addresses the reputational consequences of punishment that does not entail spite (incurring a cost to inflict a cost).

4.2.1 Study 3a Methods

4.2.1.1 Participants

Participants were 246 English speakers in the United States (48.78% female, M_{age} = 29, SD_{age} = 9) recruited via Prolific. They received 1.28 dollars for their participation, which lasted about 8 minutes. Those who wished to participate in the study first completed a written informed consent form. Studies 3a and 3b were approved by the Institutional Review Board at University of California, Santa Barbara (Human Subjects Committee).

[6] Assuming punishment is costly may stem from the intuition that defectors can retaliate against a sanctioner in real life. But that can happen to a withdrawer as well as a punisher; trade wars between nations are an example.

4.2.1.2 Design

Participants were instructed that they would observe two individuals repeatedly interact in a Dictator Game with Taking Option (DGwT) (List, 2007). It was explained that there are two roles: giver and receiver. Both individuals are given $5 at the beginning of a round; then the giver receives an additional endowment of $5. The giver decides either to share this endowment with the receiver (up to $5) or to take money from the receiver (up to $5), both in $1 increments. After the giver's decision, the two switch roles and interact again.

After the explanation, participants observed two individuals, Alex and Casey, play three rounds of DGwT. (These names were chosen because they can apply to any gender; in reporting results, both will be referred to as "she" for ease of exposition). Participants were told that Alex and Casey knew that they would interact repeatedly (participants did not know for how many rounds). In round 1, where Alex was the giver and Casey was the receiver, Alex gave $5 to Casey. In round 2, where Casey became the giver, Casey gave $0 to Alex, the receiver. Notice that Alex cooperated in round 1, and Casey failed to reciprocate in round 2.

In round 3, Alex became the giver again. Participants observed Alex make one of three responses in round 3 (between-subjects conditions):

- Punish: Alex took $5 from Casey
- Withdraw cooperation: Alex gave $0 to Casey
- No negative sanction (keep cooperating): Alex gave $5 to Casey again.

Terms such as "cooperation" and "punishment" were not used in the instructions to participants. Participants who did not understand or remember Alex's response were excluded from the study (see Appendix C).

After observing the interaction, participants evaluated Alex on 24 adjectives: exploitable, weak, gullible, unwise, incompetent, vengeful, aggressive, impulsive, cowardly, frightened, mean, careless, dependable, likable, forgiving, generous, considerate, cooperative, trustworthy, honorable, friendly, kind, fair, and emotionally-stable. The order of the adjectives was randomized. Adjectives were taken from previous research (Barclay, 2006; Delton et al., 2012; Kiyonari & Barclay, 2008; Nelissen, 2008) or unanimously nominated by the authors. Each adjective was rated on a 7-point Likert scale (from 1: "Not at all" to 7: "Extremely"). Participants also rated how much they would like to interact with Alex in a DGwT on a 5-point Likert scale (from 1: "Not at all" to 5: "Extremely").

4.2.2 Study 3a Results

4.2.2.1 Summary reputations

Data were analyzed using R 4.0.3 (R Core Team, 2020). First, we created summary reputations by using factor analysis to group related adjective ratings. Three factors were obtained on 24 adjective ratings, using the factanal function in R (R Core Team, 2020) with promax rotation, explaining 53.3% of the total variance. The number of factors was corroborated by parallel analysis using the fa.parallel function in the R package psych (Revelle, 2021).

We obtained three summary reputations by averaging the adjective ratings for each factor. Eleven adjectives, such as cooperative, trustworthy, considerate, and generous composed a summary reputation for being *cooperative* (Cronbach's $\alpha = .93$) (see table 3.1 for other adjectives and factor loadings). Four adjectives—vengeful, aggressive, mean, and forgiving (reverse-coded)—composed a summary reputation for being *vengeful* ($\alpha = .83$)

and nine adjectives, such as exploitable, gullible, weak, and unwise composed a summary reputation for being *exploitable* ($\alpha = .87$). The summary reputation for being *cooperative* was negatively correlated with the two others: $r\ (244) = -.65$, $p = 10^{-16}$ with *vengeful*; -.25, $p = 10^{-5}$ with *exploitable*. There was no significant correlation between the summary reputations for being *vengeful* and *exploitable* (.10, $p = .125$).

Table 3.1. Factor loadings of 24 adjectives in study 3a.

	Factor1 Cooperative	Factor 2 Exploitable	Factor 3 Vengeful
Considerate	**0.813**		
Cooperative	**0.797**		
Trustworthy	**0.794**		
Likable	**0.773**		
Kind	**0.754**		-0.228
Honorable	**0.750**	0.153	
Generous	**0.747**		-0.179
Dependable	**0.705**		
Fair	**0.682**	-0.199	0.318
Friendly	**0.673**		-0.250
Emotionally-stable	**0.446**	-0.184	-0.120
Gullible	0.168	**0.764**	
Weak		**0.764**	
Unwise		**0.749**	-0.161
Incompetent	-0.141	**0.747**	-0.124
Careless		**0.691**	
Exploitable	0.407	**0.678**	
Cowardly	-0.250	**0.677**	-0.114
Frightened		**0.478**	0.270
Impulsive		**0.354**	0.348
Vengeful		-0.143	**0.802**
Aggressive	-0.107		**0.657**
Mean	-0.346	0.116	**0.386**
Forgiving	0.257	0.294	**-0.654**

4.2.2.2 Reputational consequences

We compared the reputations of Alex as a punisher, withdrawer, and non-sanctioner by conducting one-way ANOVAs and post-hoc pairwise comparisons on three summary reputations, using the aov and TukeyHSD functions in R (R Core Team, 2020).

There were significant differences in how cooperative (F [2, 243] = 40.4, $p = 10^{-16}$) and vengeful (F [2, 243] = 139.7, $p = 10^{-16}$) participants found punishers, withdrawers, and non-sanctioners. Supporting H1, withdrawers were evaluated as more cooperative (5.25 vs. 4.60, $p = 10^{-5}$) and less vengeful (3.33 vs. 4.09, $p = 10^{-7}$) than punishers (figure 3.1a and b). People found punishers the least cooperative (less cooperative than non-sanctioners [5.83, $p < 10^{-16}$]) and the most vengeful (more vengeful than non-sanctioners [1.75, $p < 10^{-16}$]). Non-sanctioners were seen as more cooperative ($p = 10^{-5}$) and less vengeful ($p < 10^{-16}$) than withdrawers.

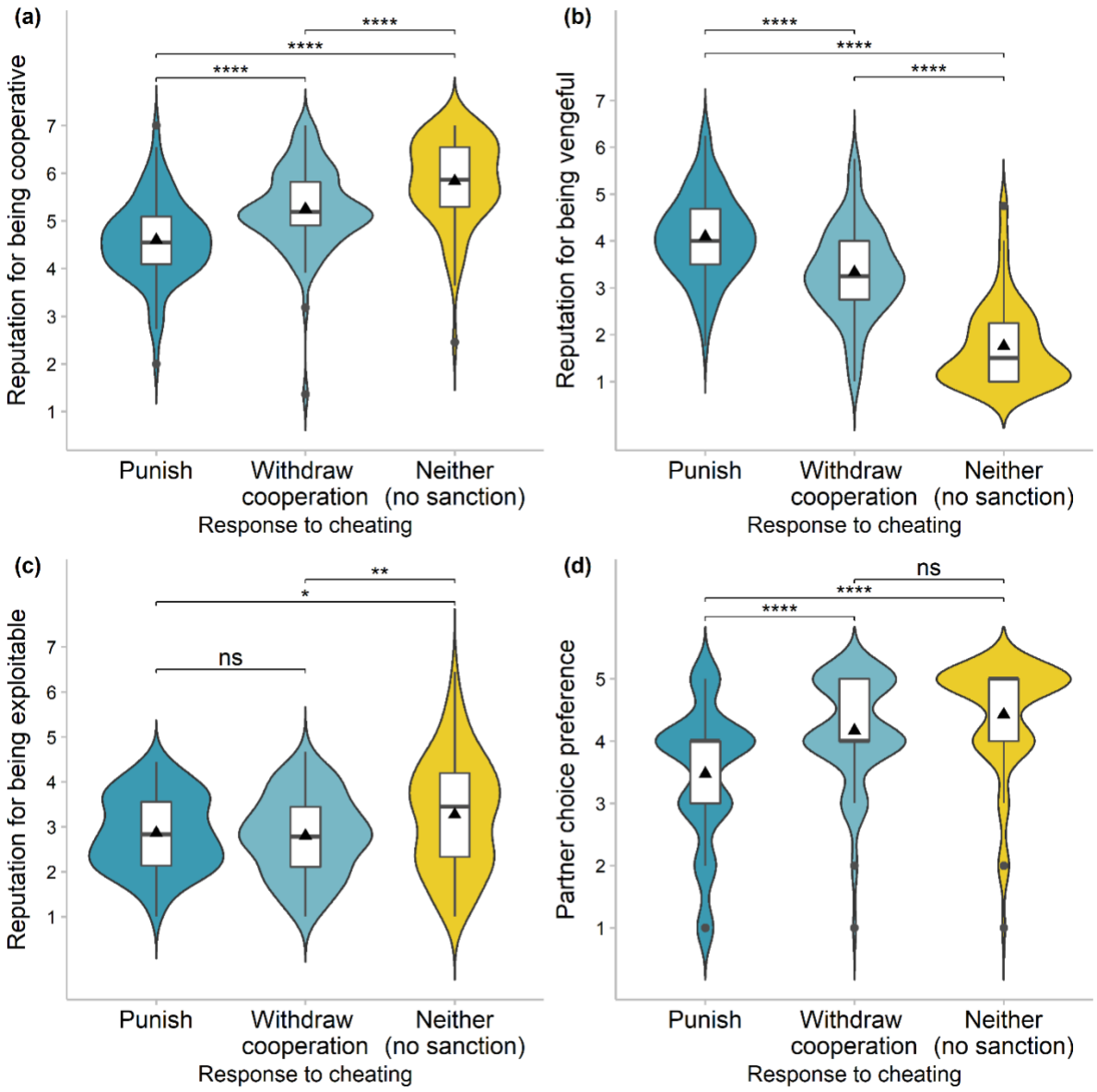


Figure 3.1. Summary reputations attributed to punishers, withdrawers, and non-sanctioners: (a) cooperative reputation, (b) vengeful reputation, and (c) exploitable reputation. (d) Partner choice preferences for each responder. Boxplots show median and quartiles; triangles represent means. *ns* > .05; * $p < .05$; ** $p < .01$; **** $p < .0001$.

There was a significant difference in how exploitable (F [2, 243] = 5.27, p = .006) participants found the three responders. People found non-sanctioners the most exploitable (figure 3.1c). Supporting H2, punishers were evaluated as less exploitable than non-

sanctioners (2.86 vs. 3.27, $p = .028$); so were withdrawers (2.80, $p = .009$). However, participants found withdrawers as unexploitable as punishers ($p = .91$).

4.2.2.3 Partner choice

Additionally, we analyzed the single-item rating of how much participants would like to interact with the three responders. There were significant differences in how desirable they were viewed as a potential cooperation partner (F [2, 243] = 22.55, $p = 10^{-9}$). Punishers were least preferred: they were rated lower than withdrawers (3.48 vs. 4.17, $p = 10^{-5}$) and non-sanctioners (4.43, $p = 10^{-9}$) (figure 3.1d). But preferences were the same for withdrawers and non-sanctioners ($p = .19$).

The partner choice preference was positively correlated with the summary reputation for being cooperative (r [244] =.74, $p = 10^{-16}$) and negatively with the ones for being vengeful (-.49, $p = 10^{-16}$) and exploitable (-.24, $p = .0002$).

When controlling for other reputations and which response Alex made, only the summary reputation for being cooperative ($\beta = .69$, $p < 10^{-16}$) significantly increased how much participants wanted to interact with Alex (multiple regression using the lm function in R [R Core Team, 2020]; see table 3.2 for a full model). (The same was true when reputations were the only predictors in the model; model fit [AIC] was slightly better when responses were also included as predictors.)

Table 3.2. Factors affecting partner choice preference in study 3a.

Predictors	*b*	*SE*	95% CI	*β*	*t*	*p*
Summary reputation for being cooperative	0.70	0.06	[0.58, 0.81]	0.69	11.54	10^{-15}
Summary reputation for being vengeful	-0.02	0.06	[-0.14, 0.09]	-0.03	-0.43	.671
Summary reputation for being exploitable	-0.05	0.05	[-0.15, 0.04]	-0.05	-1.07	.284
Being a punisher (vs. non-sanctioner)	-0.06	0.17	[-0.39, 0.27]	-0.03	-0.35	.728
Being a withdrawer (vs. non-sanctioner)	0.17	0.14	[-0.11, 0.44]	0.08	1.18	.240

Note. Adjusted $R^2 = 0.55$. CI = confidence interval for *b*.

4.2.3 Study 3a Discussion

Alex's reputation for cooperativeness differed across conditions, even though she always gave generously to Casey in the first round. She was seen as least cooperative and most vengeful when she punished Casey's defection. But does this reflect the imposition of sanctions *per se* or the effect they had on Casey's final payoff?

Table 3.3 shows the final payoffs for Alex and Casey that resulted from their interaction (after round 3) in studies 3a and 3b. Casey always gained by defecting, but by different amounts depending on how Alex responded. In study 3a, Alex's reputation for cooperativeness was highest when Casey gained the most by defecting (no sanctions), intermediate when Casey profited some by defecting (cooperation withdrawn), and lowest when the defection was punished. Vengefulness also tracked Casey's payoffs: Alex was seen as most vengeful when Casey's payoff was lowest and least vengeful when it was highest.

Table 3.3. Final payoffs to the cooperator (Alex) and the defector (Casey). *

	Study 3a Punisher recoups initial loss		Study 3b Punisher pays, loss not recouped	
	Alex	Casey	Alex	Casey
Punisher	10	5	0	5
Withdrawer	5	10	5	10
Non-sanctioner	0	15	0	15

*These are payoffs *due to their interaction*; they do not count the $5 given to both parties at the beginning of each round.

These reputational consequences could also reflect Alex's final payoffs, however, because hers were anti-correlated with Casey's ($r = -1$). Indeed, Alex profited by punishing

Casey in study 3a. What would happen to Alex's reputations if punishing made her worse off than withdrawing rather than better off?

Also, why did failing to sanction lead to Alex being seen as more exploitable than punishing or withdrawing cooperation? Was it because this was the only response with a payoff of zero in study 3a, or are sanctioners seen as less exploitable regardless of their payoff from sanctioning? We address these questions in study 3b, where punishing is costly to Alex.

4.3 Study 3b

4.3.1 Study 3b Introduction

Alex's motivation to punish was ambiguous in study 3a: Was it greed or a desire to right a wrong? By punishing Casey's failure to reciprocate, Alex inflicted a cost on a defector while also reclaiming the money she had initially given to Casey. The resulting payoff to Alex—$10—was twice the payoff Alex gained when she responded by withdrawing cooperation (table 3.3). As a withdrawer, Alex kept the $5 endowment she could have given to Casey in round 3, but she did not recoup the $5 she gave to Casey in round 1.

In study 3b, we made punishment costly to Alex. Punishing still deducted $5 from Casey, but that money did not go to Alex—Alex did not recoup her initial loss by punishing. To inflict this cost in round 3, Alex had to forgo the $5 endowment she would have kept as a withdrawer. This removes greed as a possible motive for punishment.

The resulting payoffs to both parties are shown in table 3.3. The final payoffs to Casey are identical to those in study 3a. But, unlike study 3a, where Casey's payoffs were

negatively correlated with Alex's (r = -1), there was no correlation between their payoffs in study 3b (r = 0). This allows us to see whether Alex's reputations for cooperativeness and vengefulness reflect payoffs to Casey or to Alex.

If Alex's reputation for cooperativeness reflects the benefits Casey gained from interacting with Alex, then they will follow the same pattern in both studies: Alex will be seen as more cooperative the higher the payoff to Casey. But if punishing tarnished Alex's reputation for cooperativeness in study 3a because observers inferred she was motivated by greed, then her reputation for being cooperative will not suffer when she punishes in study 3b. In study 3b, Alex earns more by withdrawing cooperation than by punishing or not sanctioning.

The design of study 3b also allows us to dissociate two possible reasons that punishing gave Alex a reputation for being less exploitable than failing to sanction in study 3a. Did this inference follow from her willingness to punish *per se* or did it reflect the relative payoffs of punishing versus not sanctioning?

In both studies, the withdrawer's payoff from the interaction was positive and the no sanction payoff was zero; by contrast, punishment created a positive payoff in study 3a and a zero payoff in study 3b. If punishing *per se* leads observers to see Alex as more difficult to exploit, then punishing will result in lower exploitability ratings than failing to sanction in both studies—the inference will not hinge on whether Alex's final payoff is positive versus zero. The alternative hypothesis is that inferences about exploitability are based on Alex's final payoff, regardless of her response. If earning nothing creates a reputation for being exploitable, then Alex will be seen as equally exploitable when her payoff is zero (from

punishing *or* failing to sanction) and less exploitable when her payoff is positive (from withdrawing cooperation).

4.3.2 Study 3b Methods

4.3.2.1 Participants

Participants were 203 English speakers in the United States (70% female, M_{age} = 19, SD_{age} = 1) recruited from an undergraduate psychology subject pool at University of California, Santa Barbara. Those who wished to participate in the study first completed a written informed consent form. The online study lasted about 10 minutes, and participants received a course credit for their participation.

4.3.2.2 Design

The design was identical to study 3a with two exceptions in how the giver and the receiver interacted. (i) Punishment was costly (i.e., the interaction was a Dictator Game with *Reducing* Option rather than a Dictator Game with *Taking* Option). The giver had to pay $5 to reduce the receiver's earnings by $5, instead of doing this by *taking* $5 from the receiver. (ii) Instructions about the giver's options were simplified: Giving (and reducing) was all or none (no $1 increments). Givers therefore had three options in study 3b: (a) give the receiver $5, (b) give the receiver $0, or (c) pay $5 to reduce the receiver's earnings by $5.

As in study 3a, Alex gave $5 in round 1 and Casey gave $0 in round 2. In round 3, participants observed Alex respond in one of three ways (between-subjects conditions):

- Punish: Alex paid $5 to reduce Casey's earnings by $5
- Withdraw cooperation: Alex gave $0 to Casey

- No negative sanction (keep cooperating): Alex gave $5 to Casey again.

In both studies, Alex punished by reducing Casey's earnings by $5. In study 3b, Alex had to pay $5 to accomplish this; in study 3a Alex accomplished the same reduction by taking $5 from Casey (see Appendix C).

4.3.3 Study 3b Results

4.3.3.1 Summary reputations

The same analysis strategy as in study 3a was used. The factor analysis revealed a very similar three factor structure, explaining 49.5% of the total variance. Of 24 adjectives, 21 loaded on the same factors in study 3b so, for ease of comparison, we will use the same labels for summary representations across both studies. (See table 3.4 for other adjectives and factor loadings.) Nine adjectives, such as generous, kind, considerate, and cooperative, composed a summary reputation for being *cooperative* (Cronbach's $\alpha = .92$) Five adjectives—vengeful, aggressive, impulsive, mean, and (un)forgiving—composed a summary reputation for being *vengeful* ($\alpha = .84$). Ten adjectives, such as incompetent, unwise, exploitable, weak, and gullible, composed a summary reputation for being *exploitable* ($\alpha = .83$). The summary reputation for being *cooperative* was negatively correlated with the two others: $r\ (201) = -.63$, $p = 10^{-16}$ with *vengeful*; -.18, $p = .013$ with *exploitable*. The correlation between the summary reputations for *vengeful* and *exploitable* was positive in both studies, but significant only in study 3b (.20, $p = .005$).

Table 3.4. Factor loadings of 24 adjectives in study 3b.

	Factor1 Cooperative	Factor 2 Exploitable	Factor 3 Vengeful
Generous	**0.825**		

Kind	**0.808**		
Considerate	**0.785**		-0.109
Friendly	**0.774**		-0.128
Likable	**0.755**	-0.103	
Dependable	**0.747**		0.136
Trustworthy	**0.699**		
Honorable	**0.676**		
Cooperative	**0.628**		
Incompetent		**0.734**	
Unwise		**0.679**	
Weak	0.108	**0.667**	
Cowardly	-0.259	**0.651**	-0.255
Careless		**0.619**	
Gullible	0.298	**0.614**	
Exploitable	0.188	**0.568**	-0.101
Frightened	0.123	**0.477**	0.304
Fair	0.296	**-0.411**	0.325
Emotionally-stable	0.306	**-0.331**	-0.176
Vengeful	-0.112	-0.173	**0.863**
Aggressive	-0.191		**0.693**
Impulsive		0.344	**0.485**
Mean	-0.296	0.275	**0.422**
Forgiving	0.350	0.322	**-0.606**

4.3.3.2 Reputational consequences

There were significant differences in how cooperative (F [2, 200] = 37.56, $p = 10^{-14}$) and vengeful (F [2, 200] = 120.4, $p = 10^{-16}$) participants found costly punishers, withdrawers, and non-sanctioners. As in study 3a, these reputations tracked the final payoffs to Casey. Alex's reputation for cooperativeness was highest when Casey gained the most by defecting (no sanctions), intermediate when Casey profited some by defecting (cooperation withdrawn), and lowest when the defection was punished (all differences significant; see figure 3.2a). Alex was seen as least vengeful when Casey's payoff was highest (no sanctions), intermediate when Casey profited some by defecting (cooperation withdrawn),

and most vengeful when Casey's payoff was lowest (punished; all differences significant; see figure 3.2b).

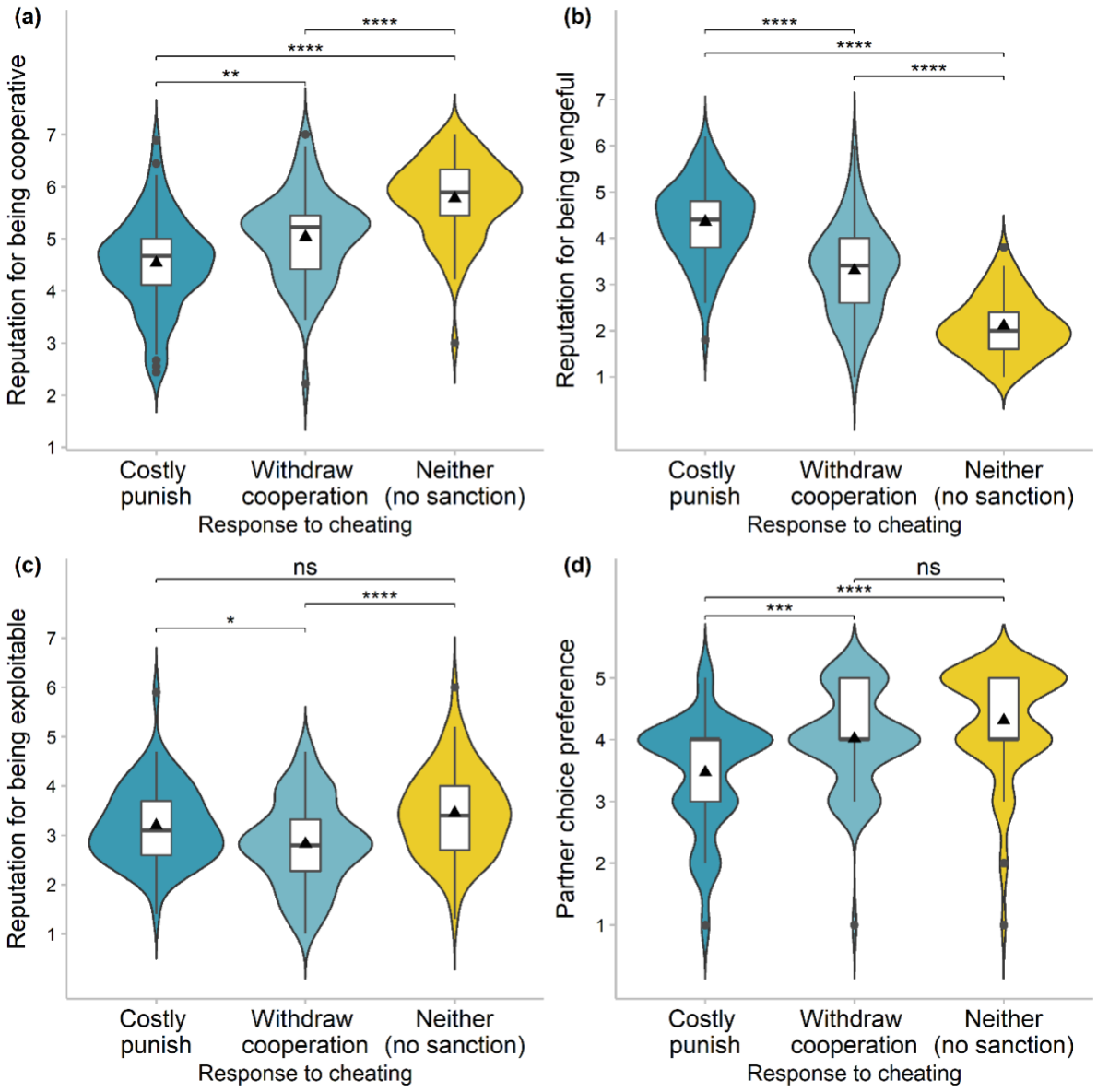


Figure 3.2. Summary reputations attributed to costly punishers, withdrawers, and non-sanctioners: (a) cooperative reputation, (b) vengeful reputation, and (c) exploitable reputation. (d) Partner choice preferences for each responder. Boxplots show median and quartiles; triangles represent means. *ns* > .05; * $p < .05$; ** $p < .01$; *** $p < .001$; **** $p < .0001$.

Casey's final payoffs were uncorrelated with Alex's in study 3b: Alex's reputations for being cooperative and vengeful *did not track Alex's final payoffs*—only Casey's.

Our main results rest on the summary representations, but curious readers can consult table 3.5 for a snapshot of how people saw costly versus non-costly punishers; it compares their ratings for each adjective in studies 3a and 3b. None of the 9 adjectives contributing to the summary representations for cooperativeness in study 3b differed for the two types of punishers. When punishing Casey's defection was costly, Alex was seen as more vengeful and impulsive than when she recouped her lost investment by punishing.

Table 3.5. Adjective ratings for punishers in studies 3a (*n* = 82) and 3b (*n* = 66).

	Study 3a: Punishers		Study 3b: Costly punishers			
	Mean	SD	Mean	SD	*t*	*p* 2-tailed (uncorrected)
Exploitable	3.48	1.45	3.94	1.33	-2.02	.045
Weak	2.66	1.24	2.67	1.14	-0.05	.960
Gullible	3.21	1.39	3.59	1.36	-1.69	.093
Unwise	2.78	1.30	3.30	1.40	-2.33	.021
Incompetent	2.37	1.20	2.79	1.23	-2.09	.038
Vengeful	4.62	1.50	5.11	1.29	-2.11	.036
Aggressive	3.79	1.42	3.92	1.28	-0.59	.555
Impulsive*	3.94	1.30	4.56	1.34	-2.85	.005
Cowardly	2.29	1.20	2.55	1.13	-1.32	.190
Frightened	2.23	1.27	2.95	1.41	-3.24	.002
Mean	2.98	1.39	3.24	1.22	-1.25	.215
Careless	2.82	1.20	3.26	1.14	-2.28	.024
Dependable	4.48	1.18	4.38	1.33	0.46	.645
Likable	4.72	1.19	4.64	1.05	0.45	.652
Forgiving	3.01	1.06	3.05	1.21	-0.18	.861
Generous	4.85	1.19	4.80	1.18	0.26	.796
Considerate	4.52	1.18	4.45	1.10	0.37	.710
Cooperative	4.67	1.08	4.61	1.15	0.35	.727
Trustworthy	4.43	1.14	4.39	1.24	0.17	.868
Honorable	4.39	1.24	4.21	1.26	0.86	.391
Friendly	4.82	1.07	4.74	1.11	0.41	.680
Kind	4.67	1.07	4.59	1.12	0.44	.661

Fair*	4.89	1.28	5.14	1.19	-1.21	.228
Emotionally-stable*	4.18	0.90	3.89	1.02	1.80	.075

Traits associated with *cooperative* are in black; those associated with *vengeful* are in red; and *exploitable* traits are in blue. *Fair and emotionally-stable loaded on *cooperative* in study 3a, but (un)fair and (un)stable loaded on *exploitable* in study 3b. *Impulsive loaded on *exploitable* in study 3a, but on *vengeful* in study 3b.
Note: The *p*-values (two-tailed) are not adjusted for multiple comparisons. Only *frightened* survives an adjustment for multiple comparisons ($p = .048$), whether using the Hommel or Benjamini-Hochberg method.

There was also a significant difference in how exploitable participants found the three responders (F [2, 200] = 9.49, $p = .0001$). The withdrawer was seen as least exploitable, with ratings lower than for the non-sanctioner (2.83 vs. 3.45, $p = 10^{-5}$) and the punisher (3.20, $p = .030$). But the exploitability of the punisher and non-sanctioner were similar ($p = .199$).

When classified by the type of response (figure 3.2c), the pattern is different from that in study 3a: Withdrawing cooperation was the only sanction that made Alex seem less exploitable in study 3b, whereas both sanctions—withdrawing and punishing—had this effect in study 3a. But when classified by Alex's final payoff due to the interaction, the results are identical across studies. Whether Alex punished or failed to sanction, a final payoff of zero led to Alex being seen as more exploitable than a final payoff that is positive. In study 3a, not sanctioning was the only response with a zero payoff for Alex; punishing and withdrawing cooperation both gave Alex a positive payoff. In study 3b, punishing and not sanctioning both led to a zero payoff for Alex; only withdrawing gave her a positive payoff.

Those curious about how costly punishers were seen compared to punishers who recouped their investment can consult table 3.5 for ratings of each adjective that loaded on

exploitability in study 3b. The snapshot for exploitability is quite different from that for cooperativeness, where none of the 9 adjectives differed in studies 3a and 3b. Costly punishers were seen as more exploitable, unwise, incompetent, frightened, and careless than punishers who recouped their loss; they also trended toward being seen as more gullible and emotionally-unstable. *N.B.* Most of these differences were not significant when corrected for multiple (24) comparisons (Benjamini & Hochberg, 1995; Hommel, 1988).

4.3.3.3 Partner choice

There were significant differences in how much participants would like to interact with the three responders (F [2, 200] = 16.76, $p = 10^{-7}$). Costly punishers were least preferred: They were rated lower than withdrawers (3.47 vs. 4.01, p = .0007) and non-sanctioners (4.30, $p = 10^{-7}$) (figure 3.2d). But preferences were similar for withdrawers and non-sanctioners (p = .12).

The partner choice preference was positively correlated with the summary reputation for being cooperative (r [201] =.72, $p = 10^{-16}$) and negatively with the ones for being vengeful (-.44, $p = 10^{-11}$) and exploitable (-.21, p = .003). When controlling for other reputations and which response Alex made, only the summary reputation for being cooperative (β = .73, $p < 10^{-16}$) significantly increased how much participants wanted to interact with Alex. (See table 3.6 for a full model.)

Table 3.6. Factors affecting partner choice preference in study 3b.

Predictors	*b*	*SE*	95% CI	*β*	*t*	*p*
Summary reputation for being cooperative	0.68	0.06	[0.56, 0.80]	0.73	11.44	10^{-15}
Summary reputation for being vengeful	-0.08	0.06	[-0.20, 0.04]	-0.08	-1.37	.171
Summary reputation for being exploitable	0.05	0.06	[-0.07, 0.18]	0.07	0.84	.400
Being a costly punisher (vs. non-sanctioner)	-0.13	0.18	[-0.49, 0.22]	-0.07	-0.74	.460
Being a withdrawer (vs. non-sanctioner)	0.10	0.14	[-0.19, 0.38]	0.05	0.68	.498

Note. Adjusted $R^2 = 0.52$. CI = confidence interval for b.

4.4 Study 3 General discussion and conclusions

There are two ways of negatively sanctioning a defector: by withdrawing cooperation or by punishing (inflicting costs). We tested whether responding to a defector by withdrawing cooperation has better reputational consequences than responding by inflicting punishment. In every condition, Alex began by cooperating generously with Casey, who failed to reciprocate this generosity. But Alex's reputation varied across conditions, depending on how she sanctioned Casey's defection. As predicted, observers saw Alex as more cooperative and less vengeful when she withdrew cooperation than when she punished. They also wanted her more as a cooperation partner when she was a withdrawer than a punisher. These results did not depend on whether inflicting punishment benefitted the punisher: By punishing, Alex recouped the loss caused by Casey's defection in study 3a, but not in study 3b, where punishment was costly. Alex's reputation for being more cooperative and less vengeful perfectly tracked Casey's payoffs, but were uncorrelated with Alex's payoffs.

We also inspected whether imposing negative sanctions on a defector prevents one from being seen as easily exploited. In study 3a, the punisher and withdrawer were both evaluated as less exploitable than the responder who imposed no negative sanctions on a defecting partner. But the punisher was not seen as more difficult to exploit than the withdrawer. In study 3b, the withdrawer was seen as less exploitable than both the punisher and non-

sanctioner. The punisher and non-sanctioner had similar reputations for exploitability when punishment was costly.

These patterns suggest that a reputation for being difficult to exploit is inferred from a sanctioner's payoffs, rather than from punishment per se. When classified based response type, the exploitability results look different for studies 3a and 3b. But they are identical when classified by whether Alex's payoff from interacting with Casey was positive versus zero. In both studies, withdrawing cooperation led to a positive payoff for Alex and a failure to sanction led to a payoff of zero; accordingly, withdrawers were seen as less exploitable than non-sanctioners in both studies. But Alex's payoffs from punishing were different in studies 3a and 3b. In study 3a, withdrawing and punishing both gave Alex a positive payoff, and both responses earned Alex a reputation as more difficult to exploit than a failure to sanction—the only response with a payoff of zero for Alex. But in study 3b, punishing at a personal cost resulted in a zero payoff to Alex, just like a failure to sanction. In this case, the withdrawer (positive payoff) was seen as less exploitable than both responders with payoffs of zero—the punisher and non-sanctioner. That is, a positive payoff always led to Alex being seen as more difficult to exploit than a payoff of zero, regardless of how Alex responded to defection. How big the positive payoff was did not seem to matter—just that it was positive rather than zero.

We found no evidence that punishing enhances one's reputation for cooperativeness. But not imposing negative sanctions on cheaters may be costly: It risks acquiring a reputation for being easy to exploit, which may attract cheaters. This finding supports the hypothesis that motivations to negatively sanction cheaters—whether by punishing or withdrawing cooperation—evolved to prevent losses by deterring mistreatment by the defector and other

observers (Delton & Krasnow, 2017; Krasnow et al., 2016; Yamagishi et al., 2009). The differences between studies 3a and 3b in perceptions of exploitability deserve further study; they suggest that the reputational consequences of sanctions will deter mistreatment more effectively when they preserve a positive payoff for the sanctioner.

Surprisingly, people preferred withdrawers and non-sanctioners as cooperative partners to the same degree in both studies. Withdrawing cooperation did not decrease desirability as a partner, but punishing did (see also Arai et al., 2022). This result highlights an advantage of withdrawing cooperation over punishment as a negative sanction: It promotes a reputation that is likely to deter exploitation while remaining favorable as a cooperative partner.

If withdrawing cooperation is better than punishment, why do people ever punish defectors? First, the benefits of being recognized as a punisher might exceed its costs in some social ecologies. When stealing resources is common in the local social ecology, as is often the case among pastoralists, acquiring a reputation for being vengeful may deter mistreatment (Cohen & Nisbett, 1996; Herrmann et al., 2008). Second, punishers may achieve a competitive advantage over others, which over-rides the reputational costs of punishing (Raihani & Bshary, 2019).

Third, not all cooperative contexts have the same incentive structure; there are situations in which withdrawing cooperation is not possible (e.g., third party punishment games) or has disadvantages over punishing (e.g., public goods games). Most studies of the reputational consequences of punishment used these situations, and compared costly punishment to not sanctioning. In third party punishment games, withdrawing cooperation is not an option for the third party, who has no opportunity to engage in cooperation with the defector; in these games, punishers were sometimes evaluated more favorably than non-sanctioners (Jordan et

al., 2016; Nelissen, 2008; Raihani & Bshary, 2015b). Punishment in public goods games has elicited mixed results (Barclay, 2006; Kiyonari & Barclay, 2008; Mifune et al., 2020). In these situations, it is difficult to withdraw cooperation from a free rider without simultaneously withdrawing it from other, contributing members of the group, and avoiding the free rider by leaving the group entails abandoning the benefits of group cooperation (Tooby et al., 2006). Punishment is a way of selectively sanctioning a free rider without losing the benefits made possible by other, contributing members of the group. Indeed, agent-based simulations show that punishing defectors in group cooperation evolves easily under many ecologically realistic conditions because, when new groups form, the defector is less likely to free ride when the punisher is also present (Krasnow et al., 2015). For these reasons, the reputational consequences of punishing versus withdrawing may differ in group cooperation compared to dyadic exchange. This possibility can be tested with studies of reputation that sharply distinguish between group cooperation and dyadic exchange.

In summary, the results of two studies demonstrated that (i) those who withdraw cooperation from cheaters are evaluated more favorably as a cooperative partner than punishers and (ii) as long as the sanction preserves a positive payoff for the sanctioner, withdrawing cooperation and inflicting punishment both protect one from acquiring a reputation that may invite exploitation.

Chapter 5: General discussion

5.1 Evidence that motivational systems are designed for reputation-based partner choice

The current dissertation provides evidence that motivational systems regulating dyadic cooperation are designed for managing reputations to be chosen as a cooperation partner. Motivations to cooperate and inflict punishment were both regulated by cues of reputation-based partner choice—cues that one is in competition to be chosen as a partner. Because desirable cooperative partners will have alternative options and exert partner choice, these cues suggest that investing in one's reputations as a valuable cooperator would be advantageous: It would increase the probability of attracting (or retaining) desirable partners.

In study 1, the cue was the number of alternative options potential partners would have. This was captured by participants' estimates of how many opportunities others have to find a new partner in their local social ecology. These estimates up-regulated motivations to reciprocate cooperation and down-regulated motivations to punish: The more outside options participants thought others would have, the more they reciprocated and the less they punished. Study 1 also manipulated a cue of whether partner switching was possible in the experiment, but this situational cue itself had no major effects on motivations to cooperate or punish. Study 2 tested effects of two situational cues that one is being recognized and evaluated as a cooperation partner. The first cue was whether one was being regarded as a potential partner, indicated by group membership; the second was whether one's reputations

were trackable due to being identified (vs. anonymous). Both cues up-regulated motivations to cooperate and down-regulated motivations to punish, although the two cues interacted differently for the two motivational calibrations.

The pattern shown in these two studies is consistent with the proposed function of motivational regulations in the presence of competition to be chosen: attracting (or retaining) desirable partners. First, up-regulating motivations to cooperate and invest in one's reputation as a reliable cooperator can increase the probability of being chosen as a partner (Barclay, 2004; Barclay & Willer, 2007; Sylwester & Roberts, 2010, 2013). A result of study 1 also adds to these findings: reciprocating instead of cheating drastically increases the probability of retaining a partner who has an option to leave. Second, the current results suggest that down-regulating motivations to punish also serves as an investment in one's cooperative reputation. Study 1 showed that punishing a partner severely decreased the probability of retaining the partner. Study 3 provided more definitive evidence: Those who punished cheaters were viewed as less cooperative than those who did not and were less preferred as a partner. These findings illustrate that inflicting punishment hurts one's reputation as a cooperator and lowers the probability of attracting or retaining desirable partners. This also suggests that there is a trade-off between acquiring punitive versus cooperative reputations. It appears that the motivational calibrations found in studies 1 and 2 were a result of a balancing act of attracting desirable partners while retaining a reputation that may deter undesirable ones.

In sum, the current results demonstrate that motivational systems use cues that reputation-based partner choice may be relevant and, in response, invest in one's reputation as a valuable cooperator by up-regulating cooperation and down-regulating punishment. The

available evidence suggests that these motivational calibrations are likely to increase the probability of attracting or retaining desirable partners. The present work contributes to the growing body of literature that motivational systems regulating cooperative interactions may have evolved due to selection pressures in biological markets (Barclay, 2004; Barclay & Willer, 2007; Sylwester & Roberts, 2010, 2013).

5.2 Reputation-based partner choice may select for punitive motivations

The theory of reputation-based partner choice generates rich testable predictions about design features of motivational systems in social exchange. So far, most of the literature has been dedicated to the study of one type of reputation, a reputation as a cooperator. It has demonstrated that motivational systems are designed to value "the reputation" as a cooperator for both choosing partners and being chosen by partners (reviewed in Barclay, 2013, 2016; Manrique et al., 2021; Roberts et al., 2021). However, the flip side is that this literature has not uncovered how these systems regulate motivations to acquire *other reputations* to solve the twin problems of choosing and being chosen.

Notably, a reputation as a punisher has not yet been integrated with the literature on reputation-based partner choice (Raihani & Bshary, 2015a): Little is known about how motivations are calibrated to acquire a punitive reputation, even though it may help one deter undesirable partners such as cheaters. This is surprising considering that punishment has been the most extensively studied method to promote cooperation (e.g., Balliet et al., 2011; Fehr & Gächter, 2002; Gardner & West, 2004). Very few studies have examined

whether acquiring a reputation as a punisher *discourages undesirable cooperation partners* such as cheaters (dos Santos et al., 2013), although it has been extensively studied whether having a reputation as a punisher *attracts desirable partners*—by leading to reputations for being cooperative, trustworthy, fair, etc. (Barclay, 2006; Dhaliwal et al., 2021; Kiyonari & Barclay, 2008; Nelissen, 2008; Mifune et al. 2000) or by being preferred as a cooperation partner (Balafoutas et al., 2014; Barclay, 2006; Barclay & Raihani, 2016; Bone et al., 2016; Fehr & Rockenbach, 2003; Horita, 2010; Fehr & Rockenbach, 2003; Jordan et al., 2016; Kiyonari & Barclay, 2008; Nelissen, 2008; Ozono & Watabe, 2012; Przepiorka & Liebe, 2016; Raihani & Bshary, 2015a). Importantly, these lines of research rarely consider the possibility that motivational systems may be designed to acquire a punitive reputation. A punitive reputation in the current literature is usually thought of as a byproduct of promoting cooperation—like "accidental fame"—that may compensate the cost of inflicting punishment.

The paucity of studies may be a logical extension of the biological market theory. Naturally, punishing a partner is unnecessary and inefficient when one can exert partner switching or partner choice, which allows one to abandon an uncooperative partner and switch to a more rewarding one (Barclay, 2013; Hammerstein & Noë, 2016). Therefore, it seems almost illogical to study punishment in the framework of reputation-based partner choice. However, partner choice was not always easy or possible in the social ecologies in which human motivational systems evolved (Arai et al., 2022). The variance in the extent to which individuals can exercise partner choice may have acted as one of the selection pressures for mechanisms regulating motivations to punish. The results of study 1 indeed show that estimates of partner choice regulate motivations to punish, indicating that

punishment has been in the human behavioral repertoire in social ecologies with varying levels of partner choice.

The current results indicate that motivations to inflict punishment are a crucial part of the psychological architecture underlying reputation management. Punitive motivations were systematically down-regulated by cues of competition to be chosen, accompanied by up-regulated motivations to cooperate. This pattern suggests that reputation-based partner choice is not dictated solely by "the reputation" for being cooperative but rather may depend on multiple reputations, including a reputation for being punitive. It may be therefore fruitful to explore how other various reputations in addition to a cooperative reputation—for example, a reputation for being forgiving, trusting, uncalculating, consistent—affect partner choice and how motivations are calibrated to acquire these reputations.

Future investigation should further inspect design features of motivational systems regulating punitive motivations. If systems are designed to invest in a punitive reputation to deter undesirable partners, cues indicating the possibility of being mistreated, e.g., the prevalence of cheating and stealing in the local social ecology, should up-regulate motivations to punish (and down-regulate motivations to cooperate). Such would be a conceptual replication to offer more evidence that punitive motivational calibrations are a part of adaptations for reputation-based partner choice.

5.3 Limitations and future directions

5.3.1 A general skepticism on manipulating reputation concern by situational cues

The current study manipulated several situational cues regarding reputation-based partner choice. Some had predicted effects and affected motivations in dyadic cooperation, but some did not, raising questions about the validity of these manipulations and the robustness of their effects. In study 1, the verbal cue instructing whether partner switching was possible had no main effect on motivations to reciprocate or punish, but the partner's actual behavior did. This could be because the situation cue was unrealistic. But an alternative explanation is that punishing or under-reciprocating opportunistically, in response to temporary situational cues, damages your reputation as a valuable cooperator.

In study 2, the cue of identifiability, whether a participant's identity and reputations were trackable, consistently regulated motivations to cooperate with ingroup members, but had no effect on motivations to punish in the initial experiment (study 2a). Its saliency had to be emphasized by several means before significant effects on punitive motivations were found in the following experiments (studies 2b and 2c). In these experiments, a partner choice phase was added to make the identifiability cue more relevant to the dyadic interactions, and participants were allowed to choose a partner either only from their own group or only from a different group. This last manipulation had no effect on motivations to cooperate or punish.

The cues manipulated may have been atypical for real-life social interactions (e.g., telling someone there is no partner switching), or in conflict with other experimental

manipulations (e.g., when you have been interacting with ingroup members, it may be odd to subsequently be prevented from choosing an ingroup member as a partner). As a result, these cues might not have been appropriate inputs for motivational systems.

Indeed, there has been considerable debate and mixed evidence for the effect of situational cues on inducing reputation concern in the lab. It has been proposed that a cue of surveillance ("watching eyes") indicates the presence of observers and thereby up-regulates people's motivations to cooperate, but reviews and meta-analyses repeatedly show that it is unclear how robust and replicable the effect is (Nettle et al., 2013; Northover et al., 2017; Sparks & Barclay, 2013). This underlines the importance of replicating the current findings. Particularly, although study 2 itself provides a conceptual replication of the key finding by Yamagishi and colleagues (e.g., Yamagishi & Mifune, 2008)—that ingroup favoritism emerges only when it can improve your reputation as a cooperator with likely partners—further research is needed to resolve the mixed results (Everett et al., 2015b; Imada, 2020; Misch et al., 2021; Romano, Balliet, & Wu, 2017).

5.3.2 Withdrawing cooperation: motivations to do so and its reputational consequences

Study 3 demonstrated that withdrawing cooperation from a cheater is more advantageous than punishing in dyadic social exchange. However, the current study did not address how systems calibrate motivations to withdraw cooperation versus punish. In fact, much is still unknown about how systems regulate motivations to withdraw cooperation when one can also choose to punish (Barclay & Raihani, 2016). Yet, the present findings can help us make several testable predictions. Considering that withdrawing cooperation is less likely to harm

one's reputation as a cooperator than punishing, we can predict that cues indicating that one is in competition to be chosen (e.g., presence of potential partners) will up-regulate motivations to withdraw cooperation and down-regulate motivations to punish. On the other hand, cues of adverse nature, such as being observed by people with a disposition to cheat, may have opposite effects: down-regulating motivations to withdraw cooperation while up-regulating motivations to punish.

Nevertheless, withdrawing cooperation is not a dichotomous choice of whether to withdraw or not. One can withdraw cooperation by decreasing the amount one provides rather than abruptly stop giving at all. That is, withdrawing cooperation may be a special case of down-regulating motivations to cooperate in response to cheating. A future study could examine the effect of experiences of being cheated and clarify whether systems regulate motivations to withdraw cooperation differently from motivations to cooperate.

Another topic that requires future investigation is reputational consequences of withdrawing (or down-regulating) cooperation in groups. As discussed in study 3, reputational consequences of withdrawing cooperation could differ in group cooperation compared to dyadic cooperation (so would motivations to withdraw cooperation in groups, which may be another topic of future investigation). But no studies have so far inspected reputational consequences of withdrawing contributions to group cooperative efforts. This is striking given the plethora of studies examining reputational consequences of punishing free riders versus not sanctioning (Barclay, 2006; Horita, 2010; Kiyonari & Barclay, 2008; Mifune et al., 2020) versus rewarding cooperators (Kiyonari & Barclay, 2008; Ozono & Watabe, 2012). It may be fruitful to compare reputational consequences of punishing versus withdrawing cooperation, as withdrawing is a typical response to free riding when

punishment is not an option (e.g., Fehr & Gächter, 2000, 2002; Yamagishi, 1986). Because withdrawing cooperation from free riders entails withdrawing it from other contributing members, people may consider it as cheating or under-contributing for self-interest rather than self-protection. Those who withdraw cooperation in group may be therefore viewed as less cooperative than those who punish and less preferred as partners. Future study can help us find whether reputational consequences of withdrawing versus punishing are reversed in dyadic and group cooperation.

5.4 Conclusion

Three studies provide convergent evidence of human motivational adaptations for reputation-based partner choice. Studies 1 and 2 show that motivations to cooperate and invest in one's reputation as a valuable cooperator are up-regulated by cues that one is in competition to be chosen as a cooperation partner. It is also shown that motivations to punish and invest in a reputation that may deter cheating are down-regulated by the very same cues. Study 3 confirms the functional logic of the punitive motivational calibration: Inflicting punishment damages one's reputation as a cooperator and lowers the probability of being chosen as a partner. These results demonstrate that motivational systems are designed for managing reputations to attract and retain desirable cooperative partners. Overall, the present research adds to the growing body of evidence that competition to be chosen as a partner may have selected for motivational systems regulating cooperation and punishment in dyadic social exchange.

References

Aktipis, C. A. (2004). Know when to walk away: Contingent movement and the evolution of cooperation. *Journal of Theoretical Biology*, *231*(2), 249–260. https://doi.org/10.1016/j.jtbi.2004.06.020

Arai, S., Tooby, J., & Cosmides, L. (2022). Motivations to reciprocate cooperation and punish defection are calibrated by estimates of how easily others can switch partners. *PLOS ONE*, 0267153.

Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*(4489), 1390–1396. https://doi.org/10.1007/978-3-540-27797-2_34

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences*, *111*(45), 15924–15927. https://doi.org/10.1073/pnas.1413170111

Balduzzi, S., Rücker, G., & Schwarzer, G. (2019). How to perform a meta-analysis with R: a practical tutorial. *Evidence-Based Mental Health*, *22*, 153–160.

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, *137*(4), 594–615. https://doi.org/10.1037/a0023489

Balliet, D., Wu, J., De Dreu, C. K. W., & Dreu, C. K. W. D. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin*, *140*(6), 1556–1581. https://doi.org/10.1037/a0037737

Barclay, P. (2004). Trustworthiness and competitive altruism can also solve the "tragedy of the commons." *Evolution and Human Behavior*, *25*(4), 209–220. https://doi.org/10.1016/j.evolhumbehav.2004.04.002

Barclay, P. (2006). Reputational benefits for altruistic punishment. *Evolution and Human Behavior*, *27*(5), 325–344. https://doi.org/10.1016/j.evolhumbehav.2006.01.003

Barclay, P. (2013). Strategies for cooperation in biological markets, especially for humans. *Evolution and Human Behavior*, *34*(3), 164–175. https://doi.org/10.1016/j.evolhumbehav.2013.02.002

Barclay, P. (2015). Reputation. In D. Buss (Ed.), *Handbook of Evolutionary Psychology (2nd ed)*. (pp. 810–828). John Wiley.

Barclay, P. (2016). Biological markets and the effects of partner choice on cooperation and friendship. *Current Opinion in Psychology*, *7*, 33–38. https://doi.org/10.1016/j.copsyc.2015.07.012

Barclay, P., & Raihani, N. (2016). Partner choice versus punishment in human Prisoner's Dilemmas. *Evolution and Human Behavior*. https://doi.org/10.1016/j.evolhumbehav.2015.12.004

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1610), 749–753. https://doi.org/10.1098/rspb.2006.0209

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Batistoni, T., Barclay, P., & Raihani, N. J. (2022). Third-party punishers do not compete to be chosen as partners in an experimental game. *Proceedings of the Royal Society B: Biological Sciences*, *289*(1966), 20211773. https://doi.org/10.1098/rspb.2021.1773

Baumard, N., André, J.-B. B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, *36*(01), 59–78. https://doi.org/10.1017/S0140525X11002202

Benjamini, Y., & Hochberg, Y. (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.

Bernhard, H., Fehr, E., & Fischbacher, U. (2006). Group Affiliation and Altruistic Norm Enforcement. *American Economic Review*, *96*(2), 217–221. https://doi.org/10.1257/000282806777212594

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature*, *442*(7105), 912–915. https://doi.org/10.1038/nature04981

Binning, S. A., Rey, O., Wismer, S., Triki, Z., Glauser, G., Soares, M. C., & Bshary, R. (2017). Reputation management promotes strategic adjustment of service quality in cleaner wrasse. *Scientific Reports*, *7*(1), 8425. https://doi.org/10.1038/s41598-017-07128-5

Bliege Bird, R., & Power, E. A. (2015). Prosocial signaling and cooperation among Martu hunters. *Evolution and Human Behavior*, *36*(5), 389–397. https://doi.org/10.1016/j.evolhumbehav.2015.02.003

Bliege Bird, R., Scelza, B., Bird, D. W., & Smith, E. A. (2012). The hierarchy of virtue: Mutualism, altruism and signaling in Martu women's cooperative hunting. *Evolution*

and Human Behavior, *33*(1), 64–78. https://doi.org/10.1016/j.evolhumbehav.2011.05.007

Boero, R., Bravo, G., Castellani, M., & Squazzoni, F. (2009). Reputational cues in repeated trust games. *Journal of Socio-Economics*, *38*(6), 871–877. https://doi.org/10.1016/j.socec.2009.05.004

Bone, J. E., Wallace, B., Bshary, R., & Raihani, N. J. (2015). The effect of power asymmetries on cooperation and punishment in a prisoner's dilemma game. *PLOS ONE*, *10*(1), e0117183. https://doi.org/10.1371/journal.pone.0117183

Bone, J. E., Wallace, B., Bshary, R., & Raihani, N. J. (2016). Power asymmetries and punishment in a prisoner's dilemma with variable cooperative investment. *PLOS ONE*, *11*(5), e0155773. https://doi.org/10.1371/journal.pone.0155773

Bradley, A., Lawrence, C., & Ferguson, E. (2018). Does observability affect prosociality? *Proceedings of the Royal Society B: Biological Sciences*, *285*(1875). https://doi.org/10.1098/rspb.2018.0116

Brewer, P., & Venaik, S. (2014). The Ecological Fallacy in National Culture Research. *Organization Studies*, *35*(7), 1063–1086. https://doi.org/10.1177/0170840613517602

Bshary, R. (2002). Biting cleaner fish use altruism to deceive image–scoring client reef fish. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *269*(1505), 2087–2093. https://doi.org/10.1098/rspb.2002.2084

Bshary, R., & Grutter, A. S. (2002). Asymmetric cheating opportunities and partner control in a cleaner fish mutualism. *Animal Behaviour*, *63*(3), 547–555. https://doi.org/10.1006/anbe.2001.1937

Bshary, R., & Grutter, A. S. (2006). Image scoring and cooperation in a cleaner fish mutualism. *Nature*, *441*(7096), 975–978. https://doi.org/10.1038/nature04755

Buhrmester, D., Goldfarb, J., & Cantrell, D. (1992). Self-Presentation when Sharing with Friends and Nonfriends. *The Journal of Early Adolescence*, *12*(1), 61–79.

Bull, J. J., & Rice, W. R. (1991). Distinguishing mechanisms for the evolution of co-operation. *Journal of Theoretical Biology*, *149*(1), 63–74. https://doi.org/10.1016/S0022-5193(05)80072-4

Clutton-Brock, T. H., & Parker, G. A. (1995). Punishment in animal societies. *Nature*, *373*(6511), 209–216. https://doi.org/10.1038/373209a0

Cohen, D., & Nisbett, R. E. (1996). *Culture of honor: The psychology of violence in the South*. Westview Press Inc.

Daly, M., & Wilson, M. (1988). *Homicide: Foundations of human behavior*. Aldine de Gruyter.

Debove, S., André, J.-B., & Baumard, N. (2015). Partner choice creates fairness in humans. *Proceedings of the Royal Society B: Biological Sciences*, *282*(1808), 20150392. https://doi.org/10.1098/rspb.2015.0392

Delton, A. W., Cosmides, L., Guemo, M., Robertson, T. E., & Tooby, J. (2012). The Psychosemantics of Free Riding: Dissecting the Architecture of Moral Concept. *Journal of Personality and Social Psychology*, *102*(6), 1252–1270. https://doi.org/10.1037/a0027026.The

Delton, A. W., & Krasnow, M. M. (2017). The psychology of deterrence explains why group membership matters for third-party punishment. *Evolution and Human Behavior*, *38*(6), 734–743. https://doi.org/10.1016/j.evolhumbehav.2017.07.003

Delton, A. W., Krasnow, M. M., Cosmides, L., & Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences*, *108*(32), 13335–13340. https://doi.org/10.1073/pnas.1102131108

Dhaliwal, N., Patil, I., & Cushman, F. (2021). Reputational and cooperative benefits of third-party compensation. *Organizational Behavior and Human Decision Processes*, *164*(January), 27–51. https://doi.org/10.1016/j.obhdp.2021.01.003

dos Santos, M., Rankin, D. J., & Wedekind, C. (2013). Human cooperation based on punishment reputation. *Evolution*, *67*(8), 2446–2450. https://doi.org/10.1111/evo.12108

Dugatkin, L. A., & Alfieri, M. (1991). Tit-For-Tat in guppies (Poecilia reticulata): The relative nature of cooperation and defection during predator inspection. *Evolutionary Ecology*, *5*(3), 300–309. https://doi.org/10.1007/BF02214234

Dunham, Y. (2018). Mere Membership. *Trends in Cognitive Sciences*, *22*(9), 780–793. https://doi.org/10.1016/j.tics.2018.06.004

Eisenbruch, A. B., Grillot, R. L., Maestripieri, D., & Roney, J. R. (2016). Evidence of partner choice heuristics in a one-shot bargaining game. *Evolution and Human Behavior*, *37*(6), 429–439. https://doi.org/10.1016/j.evolhumbehav.2016.04.002

Eisenbruch, A. B., Grillot, R. L., & Roney, J. R. (2019). Why Be Generous? Tests of the Partner Choice and Threat Premium Models of Resource Division. *Adaptive Human Behavior and Physiology*, *5*(3), 274–296. https://doi.org/10.1007/s40750-019-00117-0

Eisenbruch, A. B., & Roney, J. R. (2017). The Skillful and the Stingy: Partner Choice Decisions and Fairness Intuitions Suggest Human Adaptation for a Biological Market of Cooperators. *Evolutionary Psychological Science*, *3*(4), 364–378. https://doi.org/10.1007/s40806-017-0107-7

Everett, J. A. C., Faber, N. S., & Crockett, M. (2015a). Preferences and beliefs in ingroup favoritism. *Frontiers in Behavioral Neuroscience*, *9*(FEB), 1–21. https://doi.org/10.3389/fnbeh.2015.00015

Everett, J. A. C., Faber, N. S., & Crockett, M. J. (2015b). The influence of social preferences and reputational concerns on intergroup prosocial behaviour in gains and losses contexts. *Royal Society Open Science*, *2*(12). https://doi.org/10.1098/rsos.150546

Fehr, E., & Gächter, S. (2000). Cooperation and Punishment in Public Goods Experiments. *American Economic Association*, *90*(4), 980–994.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(415(6868)), 137–140.

Fehr, E., & Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, *422*(6928), 137–140. https://doi.org/10.1038/nature01474

Feinberg, M., Willer, R., & Schultz, M. (2014). Gossip and Ostracism Promote Cooperation in Groups. *Psychological Science*, *25*(3). https://doi.org/10.1177/0956797613510184

Foddy, M., Platow, M., J., & Yamagishi, T. (2009). Group-Based Trust in Strangers: The Role of Stereotypes and Expectations. *Psychological Science*, *20*(4), 419–422.

Funder, D. C., & Sneed, C. D. (1993). Behavioral Manifestations of Personality: An Ecological Approach to Judgmental Accuracy. *Journal of Personality and Social Psychology*, *64*(3), 479–490.

Gardner, A., & West, S. A. (2004). Cooperation and Punishment, Especially in Humans. *The American Naturalist*, *164*(6), 753–764.

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., ..., & Aycan, Z. (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, *332*(6033), 1100–1104.

Goette, L., Huffman, D., & Meier, S. (2012). The impact of social ties on group interactions: Evidence from minimal groups and randomly assigned real groups. *American Economic Journal: Microeconomics*, *4*(1), 101–115. https://doi.org/10.1257/mic.4.1.101

Grutter, A. S., & Bshary, R. (2003). Cleaner wrasse prefer client mucus: Support for partner control mechanisms in cleaning interactions. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(suppl_2). https://doi.org/10.1098/rsbl.2003.0077

Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, *35*(1), 1–59. https://doi.org/10.1177/1056492615586597

Guo, R., Ding, J., & Wu, Z. (2020). How intergroup relation moderates group bias in Third-Party Punishment. *Acta Psychologica*, *205*(July 2019), 103055. https://doi.org/10.1016/j.actpsy.2020.103055

Gurven, M., Allen-Arave, W., Hill, K., & Hurtado, M. (2000). "It's a Wonderful Life": Signaling generosity among the Ache of Paraguay. *Evolution and Human Behavior*, *21*(4), 263–282. https://doi.org/10.1016/S1090-5138(00)00032-5

Gurven, M., Hill, K., Kaplan, H., Hurtado, A., & Lyles, R. (2000). Food transfers among Hiwi foragers of Venezuela: Tests of reciprocity. *Human Ecology*, *28*(2), 171–218. https://doi.org/10.1023/A:1007067919982

Hammerstein, P., & Noë, R. (2016). Biological trade and markets. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1687). https://doi.org/10.1098/rstb.2015.0101

Hashimoto, H., & Yamagishi, T. (2016). Duality of independence and interdependence: An adaptationist perspective. *Asian Journal of Social Psychology*, *19*, 286—297.

Hayashi, N. (1993). From tit-for-tat to out-for-tat. *Scociological Theory and Methods*, *8*(1), 19–32.

Herrmann, B., Thöni, C., & Gächter, S. (2008). Antisocial punishment across societies. *Science, March*.

Hilbe, C., & Traulsen, A. (2012). Emergence of responsible sanctions without second order free riders, antisocial punishment or spite. *Scientific Reports*, *2*, 458. https://doi.org/10.1038/srep00458

Hill, K., & Hurtado, A. M. (2009). Cooperative breeding in South American hunter–gatherers. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1674), 3863–3870. https://doi.org/10.1098/rspb.2009.1061

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, *75*(2), 383–386. https://doi.org/10.1093/biomet/75.2.383

Horita, Y. (2010). Punishers May Be Chosen as Providers But Not as Recipients. *Letters on Evolutionary Behavioral Science*, *1*(1), 6–9. https://doi.org/10.5178/lebs.2010.2

Horita, Y., & Yamagishi, T. (2007). Reciprocity and identity protection: Reasons for rejection in the ultimatum game. *Japanese Journal of Psychology, 78*, 446–451.

Horita, Y., & Yamagishi, T. (2010). Adaptive foundation of group-based reciprocity. *Shinrigaku Kenkyu*, *81*(2), 114–122. https://doi.org/10.4992/jjpsy.81.114

Imada, H. (2020). *In-group favouritism in multiple social category contexts: Extending generosity towards out-group members.*

Izquierdo, S. S., Izquierdo, L. R., & Vega-Redondo, F. (2010). The option to leave: Conditional dissociation in the evolution of cooperation. *Journal of Theoretical Biology*, *267*(1), 76–84. https://doi.org/10.1016/j.jtbi.2010.07.039

Jin, N., & Yamagishi, T. (1997). Group Heuristics in Social Dilemma. *Japanese journal of social psychology*, *12*(2), 190–198.

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473–476. https://doi.org/10.1038/nature16981

Jordan, J. J., McAuliffe, K., & Warneken, F. (2014). Development of in-group favoritism in children's third-party punishment of selfishness. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(35), 12710–12715. https://doi.org/10.1073/pnas.1402280111

Joyce, D., Kennison, J., Densmore, O., Guerin, S., Barr, S., Charles, E., & Thompson, N. S. (2006). My way or the highway: A more naturalistic model of altruism tested in an iterative prisoners' dilemma. *Jasss*, *9*(2), 79–92.

Jussim, L. (1991). Social perception and social reality: A reflection-construction model. *Psychological Review*, *98*(1), 54–73. https://doi.org/10.1037//0033-295x.98.1.54

Kamei, K. (2018). The role of visibility on third party punishment actions for the enforcement of social norms. *Economics Letters*, *171*, 193–197. https://doi.org/10.1016/j.econlet.2018.07.043

Kaplan, H., Hill, K., Cadeliña, R. V., Hayden, B., Hyndman, D. C., Preston, R. J., Smith, E. A., Stuart, D. E., & Yesner, D. R. (1985). Food Sharing Among Ache Foragers: Tests of Explanatory Hypotheses [and Comments and Reply]. *Current Anthropology*, *26*(2), 223–246.

Kelly, R. L. (2003). Colonization of New Land by Hunter-Gatherers: Expectations and Implications Based on Ethnographic Data. In M. Rockman & J. Steele (Eds.), *The Colonization of Unfamiliar Landscapes: The Archaeology of Adaptation* (pp. 44–58). Routledge.

Kelsey, C., Grossmann, T., & Vaish, A. (2018). Early reputation management: Three-year-old children are more generous following exposure to eyes. *Frontiers in Psychology*, *9*, 1–9. https://doi.org/10.3389/fpsyg.2018.00698

Kiyonari, T., & Barclay, P. (2008). Cooperation in Social Dilemmas: Free Riding May Be Thwarted by Second-Order Reward Rather Than by Punishment. *Journal of Personality and Social Psychology*, *95*(4), 826–842. https://doi.org/10.1037/a0011381

Klein, S. B., Cosmides, L., Gangi, C. E., Jackson, B., Tooby, J., & Costabile, K. A. (2009). Evolution and Episodic Memory: An Analysis and Demonstration of a Social Function of Episodic Recollection. *Social Cognition*, *27*(2), 283–319. https://doi.org/10.1521/soco.2009.27.2.283

Klein, S. B., Cosmides, L., Tooby, J., & Chance, S. (2002). Decisions and the evolution of memory: Multiple systems, multiple functions. *Psychological Review*, *109*(2), 306–329. https://doi.org/10.1037/0033-295X.109.2.306

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*(17), 2693–2710. https://doi.org/10.1002/sim.1482

Komiya, A., Ohtsubo, Y., Nakanishi, D., & Oishi, S. (2019). Gift-giving in romantic couples serves as a commitment signal: Relational mobility is associated with more frequent gift-giving. *Evolution and Human Behavior*, *40*(2), 160–166. https://doi.org/10.1016/j.evolhumbehav.2018.10.003

Krasnow, M. M., Cosmides, L., Pedersen, E. J., & Tooby, J. (2012). What Are Punishment and Reputation for? *PLOS ONE*, *7*(9), e45662. https://doi.org/10.1371/journal.pone.0045662

Krasnow, M. M., Delton, A. W., Cosmides, L., & Tooby, J. (2016). Looking Under the Hood of Third-Party Punishment Reveals Design for Personal Benefit. *Psychological Science*, *27*(3), 405–418. https://doi.org/10.1177/0956797615624469

Krasnow, M. M., Delton, A. W., Tooby, J., & Cosmides, L. (2013). Meeting now suggests we will meet again: Implications for debates on the evolution of cooperation. *Scientific Reports*, *3*, 1–8. https://doi.org/10.1038/srep01747

Kurzban, R., DeScioli, P., & O'Brien, E. (2007). Audience effects on moralistic punishment. *Evolution and Human Behavior*, *28*(2), 75–84. https://doi.org/10.1016/j.evolhumbehav.2006.06.001

Leimar, O., & Hammerstein, P. (2001). Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society B: Biological Sciences*, *268*(1468), 745–753. https://doi.org/10.1098/rspb.2000.1573

Leimgruber, K. L., Shaw, A., Santos, L. R., & Olson, K. R. (2012). Young Children Are More Generous when Others Are Aware of Their Actions. *PLoS ONE*, *7*(10). https://doi.org/10.1371/journal.pone.0048292

Li, W. J., Jiang, L. L., & Perc, M. (2021). A limited mobility of minorities facilitates cooperation in social dilemmas. *Applied Mathematics and Computation*, *391*. https://doi.org/10.1016/j.amc.2020.125705

Lim, J. (2012). Welfare Tradeoff Ratios and Emotions: Psychological Foundations of Human Reciprocity [University of California, Santa Barbara]. In *ProQuest Dissertations and Theses* (Issue March). https://search-proquest-com.proxy.library.lincoln.ac.uk/docview/1012374170?accountid=16461

List, J. A. (2007). On the interpretation of giving in Dictator Games. *Journal of Political Economy*, *115*(3), 482–493.

Manrique, H. M., Zeidler, H., Roberts, G., Barclay, P., Walker, M., Samu, F., Fariña, A., Bshary, R., & Raihani, N. (2021). The psychological foundations of reputation-based cooperation. *Philosophical Transactions of the Royal Society B*.

Martin, J. W., & Cushman, F. (2015). To punish or to leave: Distinct cognitive processes underlie partner control and partner choice behaviors. *PLOS ONE*, *10*(4), e0125193. https://doi.org/10.1371/journal.pone.0125193

Martin, J. W., Young, L., & McAuliffe, K. (2020). *The impact of group membership on punishment versus partner choice*. https://doi.org/10.1017/CBO9781107415324.004

Mendoza, S. A., Lane, S. P., & Amodio, D. M. (2014). For Members Only: Ingroup Punishment of Fairness Norm Violations in the Ultimatum Game. *Social Psychological and Personality Science*, *5*(6), 662–670. https://doi.org/10.1177/1948550614527115

Mifune, N., Hashimoto, H., & Yamagishi, T. (2010). Altruism toward in-group members as a reputation mechanism. *Evolution and Human Behavior*, *31*(2), 109–117. https://doi.org/10.1016/j.evolhumbehav.2009.09.004

Mifune, N., Li, Y., & Okuda, N. (2020). The Evaluation of Second- and Third-Party Punishers. *Letters on Evolutionary Behavioral Science*, *11*(1), 6–9. https://doi.org/10.5178/lebs.2020.72

Misch, A., Paulus, M., Dunham, Y., Misch, A., Paulus, M., & Dunham, Y. (2021). Anticipation of Future Cooperation Eliminates Minimal Ingroup Bias in Children and Adults. *Journal of Experimental Psychology*.

Molho, C., Tybur, J. M., Van Lange, P. A. M., & Balliet, D. (2020). Direct and indirect punishment of norm violations in daily life. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-17286-2

Nelissen, R. M. A. (2008). The price you pay: Cost-dependent reputation effects of altruistic punishment. *Evolution and Human Behavior*, *29*(4), 242–248. https://doi.org/10.1016/j.evolhumbehav.2008.01.001

Nettle, D., Harper, Z., Kidson, A., Stone, R., Penton-Voak, I. S., & Bateson, M. (2013). The watching eyes effect in the Dictator Game: It's not how much you give, it's being seen to give something. *Evolution and Human Behavior*, *34*(1), 35–40. https://doi.org/10.1016/j.evolhumbehav.2012.08.004

Noë, R. (1990). A veto game played by baboons: A challenge to the use of the Prisoner's Dilemma as a paradigm for reciprocity and cooperation. *Animal Behaviour*, *39*(1), 78–90.

Noë, R., & Hammerstein, P. (1994). Biological markets: Supply and demand determine the effect of partner choice in cooperation, mutualism and mating. *Behavioral Ecology and Sociobiology*, *35*(1), 1–11.

Noë, R., & Hammerstein, P. (1995). Biological Markets. *Trends in Ecology & Evolution*, *10*(8), 336–339.

Northover, S. B., Pedersen, W. C., Cohen, A. B., & Andrews, P. W. (2017). Artificial surveillance cues do not increase generosity: Two meta-analyses. *Evolution and Human Behavior*, *38*(1), 144–153. https://doi.org/10.1016/j.evolhumbehav.2016.07.001

Nowak, M. A., & Sigmund, K. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, *393*(June), 573–577.

Ockenfels, A., & Werner, P. (2014). Beliefs and ingroup favoritism. *Journal of Economic Behavior and Organization*, *108*, 453–462. https://doi.org/10.1016/j.jebo.2013.12.003

Ohtsuki, H., & Iwasa, Y. (2006). The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology*, *239*(4), 435–444. https://doi.org/10.1016/j.jtbi.2005.08.008

Oishi, S., Schug, J., Yuki, M., & Axt, J. (2015). The psychology of residential and relational mobilities. In M. Gelfand, C. Chiu, & Y. Hong (Eds.), *Handbook of Advances in Culture and Psychology* (pp. 221–272). Oxford University Press.

Ozono, H., & Watabe, M. (2012). Reputational benefit of punishment: Comparison among the punisher, rewarder, and non-sanctioner. *Letters on Evolutionary Behavioral Science*, *3*(2), 21–24. https://doi.org/10.5178/lebs.2012.22

Panchanathan, K., & Boyd, R. (2004). Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature*, *432*(November), 499–502. https://doi.org/10.1038/nature02978

Paule, R. C., & Mandel, J. (1982). Consensus Values and Weighting Factors. *Journal of Research of the National Bureau of Standards*, *87*(5), 377–385. https://doi.org/10.6028/jres.087.022

Petersen, M. B., Sell, A., Tooby, J., & Cosmides, L. (2012). To punish or repair? Evolutionary psychology and lay intuitions about modern criminal justice. *Evolution and Human Behavior*, *33*(6), 682–695. https://doi.org/10.1016/j.evolhumbehav.2012.05.003

Piazza, J. (2008). The Effects of Perceived Anonymity on Altruistic Punishment. *Evolutionary Psychology*, *6*(3), 118–124. https://doi.org/10.1016/B978-0-12-375000-6.00160-9

Pietraszewski, D. (2013). *What Is Group Psychology? Adaptations for mapping shared intentional stances* (M. B. & S. G. (Eds.), Ed.; pp. 253–257).

Pietraszewski, D. (2021). Intergroup processes: Principles from an evolutionary perspective. In P. A. M. Van Lange, E. T. Higgins, & A. W. Kruglanski (Eds.), *Social Psychology: Handbook of Basic Principles* (pp. 373–391). New York: Guilford.

Pietraszewski, D., Curry, O. S., Petersen, M. B., Cosmides, L., & Tooby, J. (2015). Constituents of political cognition: Race, party politics, and the alliance detection system. *Cognition*, *140*, 24–39. https://doi.org/10.1016/j.cognition.2015.03.007

Pinto, A., Oates, J., Grutter, A., & Bshary, R. (2011). Cleaner Wrasses Labroides dimidiatus Are More Cooperative in the Presence of an Audience. *Current Biology*, *21*(13), 1140–1144. https://doi.org/10.1016/j.cub.2011.05.021

Pollet, T. V., Tybur, J. M., Frankenhuis, W. E., & Rickard, I. J. (2014). What can cross-cultural correlations teach us about human nature? *Human Nature (Hawthorne, N.Y.)*, *25*(3), 410–429. https://doi.org/10.1007/s12110-014-9206-3

Price, M. E., Cosmides, L., & Tooby, J. (2002). Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior*, *23*(3), 203–231. https://doi.org/10.1016/S1090-5138(01)00093-9

Przepiorka, W., & Liebe, U. (2016). Generosity is a sign of trustworthiness-the punishment of selfishness is not. *Evolution and Human Behavior*, *37*(4), 255–262. https://doi.org/10.1016/j.evolhumbehav.2015.12.003

Quillien, T. (2020). Evolution of conditional and unconditional commitment. *Journal of Theoretical Biology*, *492*, 110204. https://doi.org/10.1016/j.jtbi.2020.110204

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*.

Raihani, N. J., & Bshary, R. (2015a). The reputation of punishers. *Trends in Ecology and Evolution*, *30*(2), 98–103. https://doi.org/10.1016/j.tree.2014.12.003

Raihani, N. J., & Bshary, R. (2015b). Third-party punishers are rewarded, but third-party helpers even more so. *Evolution*, *69*(4), 993–1003. https://doi.org/10.1111/evo.

Raihani, N. J., & Bshary, R. (2019). Punishment: One tool, many uses. *Evolutionary Human Sciences*, *1*, 1–26. https://doi.org/10.1017/ehs.2019.12

Revelle, W. (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research.*

Rizopoulos, D. (2022). *GLMMadaptive: Generalized Linear Mixed Models using Adaptive Gaussian Quadrature* (0.8-5) [Computer software]. https://CRAN.R-project.org/package=GLMMadaptive

Roberts, G. (1998). Competitive altruism: From reciprocity to the handicap principle. *Proceedings of the Royal Society B: Biological Sciences*, *265*(1394), 427–431. https://doi.org/10.1098/rspb.1998.0312

Roberts, G., Raihani, N. J., Bshary, R., Manrique, H. M., Fariña, A., Samu, F., & Barclay, P. (2021). The benefits of being seen to help others: Indirect reciprocity and reputation-based partner choice. *Philosophical Transactions of the Royal Society B*, *376.*

Robinson, C., & Schumacker, R. (2009). Interaction effects: Centering, variance inflation factor, and interpretation issues. *Multiple Linear Regression Viewpoints*, *35*(1), 6–11.

Romano, A., Balliet, D., & Wu, J. (2017). Unbounded indirect reciprocity: Is reputation-based cooperation bounded by group membership? *Journal of Experimental Social Psychology*, *71*, 59–67. https://doi.org/10.1016/j.jesp.2017.02.008

Romano, A., Balliet, D., Yamagishi, T., & Liu, J. H. (2017). Parochial trust and cooperation across 17 societies. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(48), 12702–12707. https://doi.org/10.1073/pnas.1712921114

Schiller, B., Baumgartner, T., & Knoch, D. (2014). Intergroup bias in third-party punishment stems from both ingroup favoritism and outgroup discrimination.

Evolution and Human Behavior, *35*(3), 169–175. https://doi.org/10.1016/j.evolhumbehav.2013.12.006

Schuessler, R. (1989). Exit Threats and Cooperation under Anonymity. *Journal of Conflict Resolution*, *33*(4), 728–749. https://doi.org/10.1177/0022002789033004007

Schug, J., Yuki, M., Horikawa, H., & Takemura, K. (2009). Similarity attraction and actually selecting similar others: How cross-societal differences in relational mobility affect interpersonal similarity in Japan and the USA. *Asian Journal of Social Psychology*, *12*(2), 95–103. https://doi.org/10.1111/j.1467-839X.2009.01277.x

Schug, J., Yuki, M., & Maddux, W. (2010). Relational Mobility Explains Between- and Within-Culture Differences in Self-Disclosure to Close Friends. *Psychological Science*, *21*(10), 1471–1478. https://doi.org/10.1177/0956797610382786

Schweinfurth, M. K., & Taborsky, M. (2020). Rats play tit-for-tat instead of integrating social experience over multiple interactions. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1918). https://doi.org/10.1098/rspb.2019.2423

Sell, A., Tooby, J., & Cosmides, L. (2009). Formidability and the logic of human anger. *Proceedings of the National Academy of Sciences*, *106*(35), 15073–15078. https://doi.org/10.1073/pnas.0904312106

Shaw, A., Montinari, N., Piovesan, M., Olson, K. R., Gino, F., & Norton, M. I. (2014). Children develop a veil of fairness. *Journal of Experimental Psychology: General*, *143*(1), 363–375. https://doi.org/10.1037/a0031247

Shinada, M., Yamagishi, T., & Ohmura, Y. (2004). False friends are worse than bitter enemies:“Altruistic” punishment of in-group members. *Evolution and Human Behavior*, *25*(6), 379–393. https://doi.org/10.1016/j.evolhumbehav.2004.08.001

Simms, E. L., Taylor, D. L., Povich, J., Shefferson, R. P., Sachs, J. L., Urbina, M., & Tausczik, Y. (2006). An empirical test of partner choice mechanisms in a wild legume-rhizobium interaction. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1582), 77–81. https://doi.org/10.1098/rspb.2005.3292

Simpson, B. (2006). Social identity and cooperation in social dilemmas. *Rationality and Society*, *18*(4), 443–470. https://doi.org/10.1177/1043463106066381

Smith, K. M., Larroucau, T., Mabulla, I. A., & Apicella, C. L. (2018). Hunter-Gatherers Maintain Assortativity in Cooperation despite High Levels of Residential Change and Mixing. *Current Biology*, 3152–3157. https://doi.org/10.1016/j.cub.2018.07.064

Sommerfeld, R. D., Krambeck, H.-J., & Milinski, M. (2008). Multiple gossip statements and their effect on reputation and trustworthiness. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1650), 2529–2536. https://doi.org/10.1098/rspb.2008.0762

Sommerfeld, R. D., Krambeck, H.-J., Semmann, D., & Milinski, M. (2007). Gossip as an alternative for direct observation in games of indirect reciprocity. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(44), 17435–17440. https://doi.org/10.1073/pnas.95.23.13755

Sparks, A., & Barclay, P. (2013). Eye images increase generosity, but not for long: The limited effect of a false cue. *Evolution and Human Behavior*, *34*(5), 317–322. https://doi.org/10.1016/j.evolhumbehav.2013.05.001

Spears, R. (2021). Social Influence and Group Identity. *Annual Review of Psychology*, *72*, 367–390. https://doi.org/10.1146/annurev-psych-070620-111818

Sperber, D., & Baumard, N. (2012). Moral Reputation: An Evolutionary and Cognitive Perspective. *Mind and Language*, *27*(5), 495–518. https://doi.org/10.1111/mila.12000

Sugiyama, L. S. (2004). Illness, Injury, and Disability among Shiwiar Forager-Horticulturalists: Implications of Health-Risk Buffering for the Evolution of Human Life History. *American Journal of Physical Anthropology*, *123*(4), 371–389. https://doi.org/10.1002/ajpa.10325

Sylwester, K., & Roberts, G. (2010). Cooperators benefit through reputation-based partner choice in economic games. *Biology Letters*, *6*(5), 659–662. https://doi.org/10.1098/rsbl.2010.0209

Sylwester, K., & Roberts, G. (2013). Reputation-based partner choice is an effective alternative to indirect reciprocity in solving social dilemmas. *Evolution and Human Behavior*, *34*(3), 201–206. https://doi.org/10.1016/j.evolhumbehav.2012.11.009

Tajfel, H. (1982). Social Psychology of Intergroup Relations. *Annual Review of Psychology*, *33*(1), 1–39. https://doi.org/10.1146/annurev.ps.33.020182.000245

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178. https://doi.org/10.1002/ejsp.2420010202

Thomson, R., Yuki, M., Talhelm, T., Schug, J., Kito, M., Ayanian, A. H., Becker, J. C., Becker, M., Chiu, C., Choi, H.-S., Ferreira, C. M., Fülöp, M., Gul, P., Houghton-Illera, A. M., Joasoo, M., Jong, J., Kavanagh, C. M., Khutkyy, D., Manzi, C., …

Visserman, M. L. (2018). Relational mobility predicts social behaviors in 39 countries and is tied to historical farming and threat. *Proceedings of the National Academy of Sciences*, *115*(29), 7521–7526. https://doi.org/10.1073/pnas.1713191115

Thorndike, E. L. (1939). On the Fallacy of Imputing the Correlations Found for Groups to the Individuals or Smaller Groups Composing Them. *The American Journal of Psychology*, *52*(1), 122. https://doi.org/10.2307/1416673

Tooby, J., & Cosmides, L. (1996). Friendship and the Banker's Paradox: Other pathways to the Evolution of Adaptations for Altruism. *Proceedings of the British Academy*, *88*(5), 119–143. https://doi.org/10.1002/(SICI)1520-6300(1998)10:5<681::AID-AJHB16>3.3.CO;2-I

Tooby, J., Cosmides, L., & Price, M. E. (2006). Cognitive adaptations for n-person exchange: The evolutionary roots of organizational behavior. *Managerial and Decision Economics*, *27*(2–3), 103–129. https://doi.org/10.1002/mde.1287

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, *46*(1), 35–57.

Valenzuela, A., & Srivastava, J. (2012). Role of Information Asymmetry and Situational Salience in Reducing Intergroup Bias: The Case of Ultimatum Games. *Personality and Social Psychology Bulletin*, *38*(12), 1671–1683. https://doi.org/10.1177/0146167212458327

Viechtbauer, W. (2005). Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261–293. https://doi.org/10.3102/10769986030003261

Wang, C. S., & Leung, A. K. -y. (2010). The Cultural Dynamics of Rewarding Honesty and Punishing Deception. *Personality and Social Psychology Bulletin*, *36*(11), 1529–1542. https://doi.org/10.1177/0146167210385921

Williams, G. (1966). *Adaptation and Natural Selection*. Princeton University Press.

Wu, J., Balliet, D., & Van Lange, P. A. M. (2015). When Does Gossip Promote Generosity? Indirect Reciprocity Under the Shadow of the Future. *Social Psychological and Personality Science*, *6*(8), 923–930. https://doi.org/10.1177/1948550615595272

Wu, J., Balliet, D., & Van Lange, P. A. M. (2016a). Gossip Versus Punishment: The Efficiency of Reputation to Promote and Maintain Cooperation. *Scientific Reports*, *6*(December 2015), 1–8. https://doi.org/10.1038/srep23919

Wu, J., Balliet, D., & Van Lange, P. A. M. (2016b). Reputation management: Why and how gossip enhances generosity. *Evolution and Human Behavior*, *37*(3), 193–201. https://doi.org/10.1016/j.evolhumbehav.2015.11.001

Wu, Z., Chen, X., Gros-Louis, J., & Su, Y. (2018). 'She is looking at me! Shall I share?' How Chinese and American preschoolers respond to eye gaze during sharing. *Social Development*, 1–14. https://doi.org/10.1111/sode.12278

Yamagishi, T. (1986). The Provision of a Sanctioning System as a Public Good predictions derived from the new approach in an experiment. *Journal Orpersonality and Social Psychology*, *51*(1), 110–116. https://doi.org/10.1037/0022-3514.51.1.110

Yamagishi, T. (2011). *Trust: The evolutionary game of mind and society*. New York: Springer.

Yamagishi, T., Hashimoto, H., & Schug, J. (2008). Preferences versus strategies as explanations for culture-specific behavior. *Psychological Science*, *19*(6), 579–584. https://doi.org/10.1111/j.1467-9280.2008.02126.x

Yamagishi, T., Hayashi, N., & Jin, N. (1994). Prisoner's dilemma networks: Selection strategy versus action strategy. In U. Schulz, W. Albers, & U. Mueller (Eds.), *Social Dilemmas and Cooperation* (pp. 233–250). Berlin: Springer-Verlag. https://doi.org/10.1007/978-3-642-78860-4

Yamagishi, T., Horita, Y., Takagishi, H., Shinada, M., Tanida, S., & Cook, K. S. (2009). The private rejection of unfair offers and emotional commitment. *Proceedings of the National Academy of Sciences*, *106*(28), 11520–11523. https://doi.org/10.1073/pnas.0900636106

Yamagishi, T., Jin, N., & Kiyonari, T. (1999). Bounded Generalized Reciprocity: Ingroup boasting and ingroup favoritism. *Advances in Group Processes*, *16*(1), 161–197.

Yamagishi, T., & Kiyonari, T. (2000). The Group as the Container of Generalized Reciprocity. *Social Psychology Quarterly*, *63*(2), 116–132.

Yamagishi, T., Makimura, Y., Foddy, M., Matsuda, M., Kiyonari, T., & Platow, M. J. (2005). Comparisons of Australians and Japanese on group-based cooperation. *Asian Journal of Social Psychology*, *8*(2), 173–190. https://doi.org/10.1111/j.1467-839x.2005.00165.x

Yamagishi, T., & Matsuda, M. (2003). *The Role of Reputation in Open and Closed Groups: An Experimental Study of Online Trading*.

Yamagishi, T., & Mifune, N. (2008). Does shared group membership promote altruism?: Fear, greed, and reputation. *Rationality and Society*, *20*(1), 5–30. https://doi.org/10.1177/1043463107085442

Yamagishi, T., & Mifune, N. (2009). Social exchange and solidarity: In-group love or out-group hate? *Evolution and Human Behavior*, *30*(4), 229–237. https://doi.org/10.1016/j.evolhumbehav.2009.02.004

Yamagishi, T., & Mifune, N. (2016). Parochial altruism: Does it explain modern human group psychology? *Current Opinion in Psychology*, *7*, 39–43. https://doi.org/10.1016/j.copsyc.2015.07.015

Yudkin, D. A., Rothmund, T., Twardawski, M., Thalla, N., & Van Bavel, J. J. (2016). Reflexive intergroup bias in third-party punishment. *Journal of Experimental Psychology: General*, *145*(11), 1448–1459. https://doi.org/10.1037/xge0000190

Yuki, M., & Schug, J. (2020). Psychological consequences of relational mobility. *Current Opinion in Psychology*, *32*, 129–132. https://doi.org/10.1016/j.copsyc.2019.07.029

Yuki, M., Schug, J., Horikawa, H., Takemura, K., Sato, K., Yokota, K., & Kamaya, K. (2007). Development of a scale to measure perceptions of relational mobility in society. *CERSS Working Paper 75, Center for Experimental Research in Social Sciences, Hokkaido University*, *Study 2*, 4–6.

Ziegler, M., Simon, M. H., Hall, I. R., Barker, S., Stringer, C., & Zahn, R. (2013). Development of Middle Stone Age innovation linked to rapid climate change. *Nature Communications*, *4*(1), 1905. https://doi.org/10.1038/ncomms2897

Appendix A

Instructions in study 1

You are going to interact with other people who are participating in this study.
Other people will be your partner(s) in various situations in which you can benefit each other.

First, we will give you **points** that can be used throughout this study.
You and your partners can give each other **points** in various situations.
Imagine **that points are something like money – e.g., the program converts points to real money at the end of the study.**

The number of points you will earn during the study *depends on both your decisions and your partners' decisions*.
There are no right or wrong decisions. We will explain how everything works before you and your partners make any decisions.
You will be given points at the beginning of each interaction.

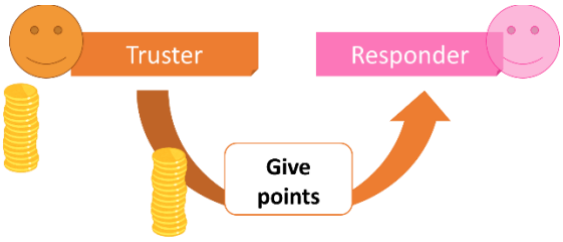
During that interaction, *you might lose points, depending on your decisions and your partners' decisions*.
However, since **we will give you points for every interaction in which you engage**, you will always have a positive number of points ***in total*** *by the end of the study*. You will never lose money by interacting with other people in this study.

Instructions for the TGP

You will be either **the truster** or **the responder**.

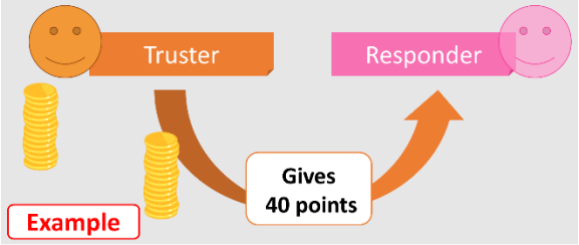
There are **four** steps.
Step 1:
The truster is given 100 points.

The truster can give **the responder** any number of these points, from 0-100.

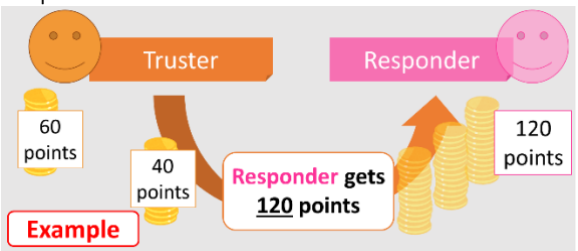


Step 2:
Whatever **the truster** gives to **the responder** is **tripled**. It becomes more valuable!

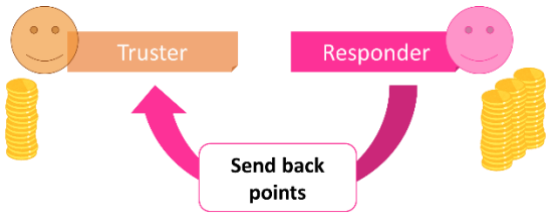
For example, suppose **the truster** gives 40 points to **the responder**.



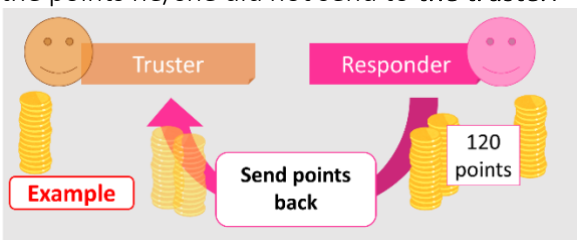
The 40 points is tripled to 120 points. So **the responder** receives 120 points, while **the truster** keeps 60 points for himself/herself.
Notice: because the points were tripled, both of you could be better off, depending on what the responder does next.



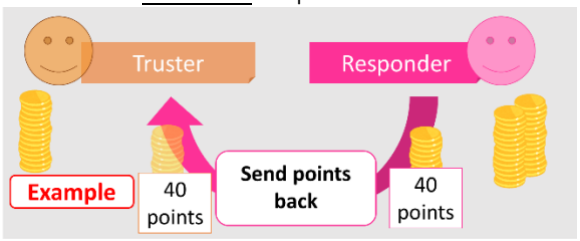
Step 3:
After receiving **the tripled points**, **the responder** can send points back to **the truster**. Any number **the responder** wants to.



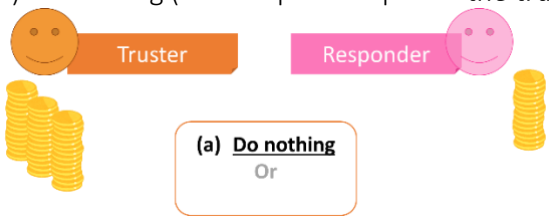
For example, if the tripled amount is 120, **the responder** can send some of those 120 points back—or none of them or all of them. As many as **the responder** wants. **The responder** keeps the points he/she did not send to **the truster**.



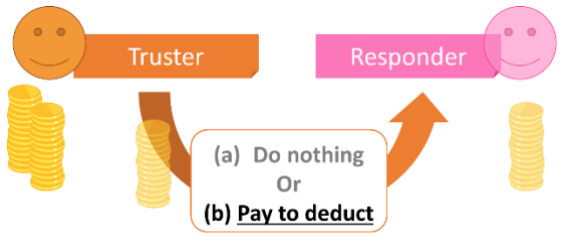
Notice: Whether **the truster** is better off than before depends **on how much of the tripled amount the responder sends back**.
For example, if **the truster** initially gave 40 points, he/she will be better off than before if **the responder** sends back more than 40 points. But the truster will be worse off if **the responder** sends back less than 40 points.



Step 4:
After receiving the points **the responder** has sent back, **the truster** can either:
(a) Do nothing (and keep all the points **the truster** now has), **OR**



(b) Pay points to deduct points from **the responder**. **Every 10 points the truster** pays deducts **30 points** from **the responder**.



Here is a summary of the four steps.

Would you like to read the explanation again?

- Yes
- No

[If Yes, go back to the top of the instruction]

Instructions for partner switching

We will pair you with a partner.
You will interact with the same partner once as the responder and once as the truster with the same partner, alternating the roles.

After interacting with the same partner once as the responder and once as the truster (this is one block),
[High Partner Choice Condition]
you will be able to switch partners (if you want to).
You will keep the same partner *only if you and your partner both choose not to switch*.

[Low Partner Choice Condition]
you will have the same partner for all other interactions.

When you are paired with a partner, the program will randomly assign you a role, either the truster or the responder. Then you and your partner will alternate the roles.

After that, the program will randomly decide whether you have a next block.

Two practice rounds of the TGP

Let's practice before you actually interact with your partner.
You will practice several blocks, but note that you do not actually win or lose points during the practice session.
After the practice session, you will be given actual points and will be able to earn more points with it.

This turn, you are the truster and your partner is the responder.
Now you and your partner are each given 50 points (a bonus for you two, regardless of your role).
…
[See "Instruction during the TGP" below for details. Participants experienced two practice rounds of the TGP (once as the truster, once as the responder) and then were asked whether they wanted to switch partners in High Partner Choice condition (or they were reminded that they have the same partner in Low Partner Choice Condition). They always practiced the truster first. The sham practice partner behaved in either of the following ways (i) returns 70%, trusts 100 points, and pays 20 points to punish if being returned less than 50%, or (ii) returns 20%, trusts 30 points, and does not punish.**]**

The program has decided that there will be no more blocks. This ends this section of your session.

This concludes the practice session.

Comprehension check questions for the TGP

Here are a few quizzes to help you understand how you interact with your partners.

Q1.
In the first step, both the truster and the responder get 50 points (as a bonus) and the truster gets 100 points more.
The truster can trust the responder and give any number of points, from 0 to 100, to the responder.

What happens to the responder if the truster gives 40 points to the responder?

- The responder receives 40 points
- The responder receives 120 points

- The responder receives **200 points**

[If choosing the correct answer, participants proceeded to the next question. Otherwise, participants were told their answer was not correct and then given a chance to answer the same question again.]
Your answer is correct! If the truster gives **40 points** to the responder, the responder receives **the tripled points: 120 points**.

[If missing the same question twice]
Your answer is **not correct**. If the truster gives **40 points** to the responder, the responder receives **the tripled points: 120 points**.

Q2.
How many points in total would **the truster** have if he/she gave **40 points** to **the responder**?

- The truster would have 60 **points** left for himself/herself
- The truster would have **110 points left**, **50 from the initial bonus**, and **the 60 that remain after the truster gave 40 points** to the receiver.
- The truster has **40 points** left for himself/herself

[If choosing the correct answer]
Your answer is correct!
At first, the truster had 50 points (a bonus). Then, the truster was given 100 points more and asked to decide how many points they would like to give to the responder *from that 100 points*. If the truster gave 40 points to the responder, the truster would **keep 60 points out of 100 points**, but the truster **still has the original 50 points**. So the truster has 60+50 = 110.

[If missing the same question twice]
Your answer is **not correct**.
At first, the truster had 50 points (a bonus). Then, the truster was given 100 points more and asked to decide how many points they would like to give to the responder *from that 100 points*. If the truster gave 40 points to the responder, the truster would **keep 60 points out of 100 points**, but the truster **still has the original 50 points**. So the truster has 60+50 = 110.

Q3.
In the **third** step, the **responder** divides the tripled points between him/herself and **the truster**.

What happens to the truster if the responder gives **100 points**?

- The truster receives **300 points**
- The truster receives **200 points**
- The truster receives **100 points**

[If choosing the correct answer]
Your answer is correct! If the **responder** sends **100 points** back to the responder, **the truster receives what is sent by the responder**: **100 points.**

[If missing the same question twice]
Your answer is **not correct**. If the **responder** sends **100 points** back to the responder, **the truster receives what is sent by the responder**: **100 points.**

Q4.
After receiving points from the responder, the truster can either (a) do nothing (pay 0 points) or (b) **pay points to deduct points from the responder**.

What happens to the responder if the truster pays **10 points**?

- The responder loses **50 points**
- The responder loses **30 points**
- The responder loses **100 points**

[If choosing the correct answer]
Your answer is correct! If the truster **pays 10 points**, the responder loses **30 points. Every 10 points** the truster pays **deducts 30 points** from the responder.

[If missing the same question twice]
Your answer is **not correct**. If the truster **pays 10 points**, the responder loses **30 points. Every 10 points** the truster pays **deducts 30 points** from the responder.

Q5a. [Only for those in Low Partner Choice Condition]
After interacting with a partner for a block (both as the giver and the responder), are you going to **have the same partner**?

- Yes
- No

[If choosing the correct answer]
Your answer is correct! After the first block, **you will have the same partner** in the following blocks.

[If missing the same question twice]
Your answer is **not correct**. After the first block, **you will have the same partner** in the following blocks.

Q5b. [Only for those in High Partner Choice Condition]
After interacting with a partner for a block (both as the giver and the responder), are you going to **be able to switch to a new partner**?

- Yes
- No

[If choosing the correct answer]

Your answer is correct! After the first block, **you will be able to switch partners** in the following blocks. You will keep the same partner *only if you and your partner both choose not to switch*.

[If missing the same question twice]
Your answer is **not correct**. After the first block, **you will be able to switch partners** in the following blocks. You will keep the same partner *only if you and your partner both choose not to switch*.

Instruction before the TGP

Now you are going to **actually interact with your partner**.
Remember that the program **will randomly decide whether you have a next block** at the end of each block.

[High Partner Choice Condition]
Remember, after the first block, you will be **able to switch partners**, if you want to, for the following blocks. After each block, you will have the option to switch partners if you want to.

[Low Partner Choice Condition]
Remember, after the first block, you will **have the same partner** for all of the following blocks.

Instruction during the TGP

Your partner will know you as Participant **R[random two-digit number].**

Your partner is **Participant S[random two-digit number]**.

[Counter-balance: If participants became the truster]

This turn, you are **the truster** and your partner is the responder.
Now you and your partner are each given **50 points** (a bonus for you two, regardless of your role).

Truster	*Responder*
[Participant's ID] (you)	**[Sham partner's ID] (your partner)**
50 points	50 points

Truster	*Responder*
[Participant's ID] (you)	**[Sham partner's ID] (your partner)**
50 points	50 points
[Participant's ID], we are giving you **100 points** more.	

How many points do you want to give to your partner? Whatever you give will be **tripled.**	

How many points do you want to give to your partner?

- 0 points, which will become 0 points for your partner
- 10 points, which will become 30 points for your partner
- 20 points, which will become 60 points for your partner
- 30 points, which will become 90 points for your partner
- 40 points, which will become 120 points for your partner
- 50 points, which will become 150 points for your partner
- 60 points, which will become 180 points for your partner
- 70 points, which will become 210 points for your partner
- 80 points, which will become 240 points for your partner
- 90 points, which will become 270 points for your partner
- 100 points, which will become 300 points for your partner

Truster	*Responder*
[Participant's ID] (you)	**[Sham partner's ID] (your partner)**
50 points	50 points
[Participant's ID], we are giving you **100 points** more. How many points do you want to give to your partner? Whatever you give will be **tripled.**	
[Participant's ID], you chose to give **XX points**	Which became **3*XX points**

Truster	*Responder*
[Participant's ID] (you)	**[Sham partner's ID] (your partner)**
50 points	50 points
[Participant's ID], we are giving you **100 points** more. How many points do you want to give to your partner? Whatever you give will be **tripled.**	
[Participant's ID], you chose to give **XX points**	Which became **3*XX points**

[Participant's ID], you now have: 150 - XX points	[Sham partner's ID], now you have: 3*XX + 50 points

Truster	*Responder*
[Participant's ID] (you)	[Sham partner's ID] (your partner)
50 points	50 points
[Participant's ID], we are giving you 100 points more. How many points do you want to give to your partner? Whatever you give will be tripled.	
[Participant's ID], you chose to give XX points	Which became 3*XX points
[Participant's ID], you now have: 150 - XX points	[Sham partner's ID], now you have: 3*XX + 50 points
	[Sham partner's ID], you can now send points back to [Participant's ID] from what he/she gave you: 3*XX points. How many points do you want to send back to your partner?

Truster	*Responder*
[Participant's ID] (you)	[Sham partner's ID] (your partner)
50 points	50 points
[Participant's ID], we are giving you 100 points more. How many points do you want to give to your partner? Whatever you give will be tripled.	
[Participant's ID], you chose to give XX points	Which became 3*XX points
[Participant's ID], you now have: 150 - XX points	[Sham partner's ID], now you have: 3*XX + 50 points
	[Sham partner's ID], you can now send points back to [Participant's ID] from what he/she gave you: 3*XX

	points. How many points do you want to send back to your partner?
[Participant's ID], your partner sent back **XXX points,** **X%** of the tripled amount that you gave	[Sham partner's ID], you sent back **XXX points**, **X%** of what [Participant's ID] gave to you. You kept **XXX points**, **X%** of it for yourself

[Sham partner sends back either <u>50% or 20%</u>]

Truster	*Responder*
[Participant's ID] (you)	**[Sham partner's ID] (your partner)**
50 points	50 points
[Participant's ID], we are giving you **100 points** more. How many points do you want to give to your partner? Whatever you give will be **tripled.**	
[Participant's ID], you chose to give **XX points**	Which became **3*XX points**
[Participant's ID], you now have: **150 - XX points**	[Sham partner's ID], now you have: **3*XX + 50 points**
	[Sham partner's ID], you can now send points back to [Participant's ID] from what he/she gave you: **3*XX points**. How many points do you want to send back to your partner?
[Participant's ID], your partner sent back **XXX points**, **X%** of the tripled amount that you gave	[Sham partner's ID], you sent back **XXX points**, **X%** of what [Participant's ID] gave to you. You kept **XXX points**, **X%** of it for yourself
[Participant's ID], you now have **XXXX points**	[Sham partner's ID], now you have **XXXX points**

What would you like to do?

- Do nothing (pay 0 points)

- Pay 10 points to deduct 30 points from your partner
- Pay 20 points to deduct 60 points from your partner
- Pay 30 points to deduct 90 points from your partner
- Pay 40 points to deduct 120 points from your partner
- Pay 50 points to deduct 150 points from your partner

Truster	*Responder*
[Participant's ID] (you)	[Sham partner's ID] (your partner)
50 points	50 points
[Participant's ID], we are giving you **100 points** more. How many points do you want to give to your partner? Whatever you give will be **tripled.**	
[Participant's ID], you chose to give **XX points**	Which became **3*XX points**
[Participant's ID], you now have: **150 - XX points**	[Sham partner's ID], now you have: **3*XX + 50 points**
	[Sham partner's ID], you can now send points back to [Participant's ID] from what he/she gave you: **3*XX points**. How many points do you want to send back to your partner?
[Participant's ID], your partner sent back **XXX points**, **X%** of the tripled amount that you gave	[Sham partner's ID], you sent back **XXX points**, **X%** of what [Participant's ID] gave to you. You kept **XXX points**, **X%** of it for yourself
[Participant's ID], you now have **XX points**	[Sham partner's ID], now you have **XX points**
[Participant's ID], do you want to deduct points from [Sham partner's ID]? **Every 10 points** you pay deduct **30 points** of [Sham partner's ID]	
[Participant's ID], you chose to pay **XX points**	[Sham partner's ID], your partner deducted **XXX points** from you

to deduct **XX points** from [Sham partner's ID]	

Truster	*Responder*
[Participant's ID] (you)	**[Sham partner's ID] (your partner)**
50 points	50 points
[Participant's ID], we are giving you **100 points** more. How many points do you want to give to your partner? Whatever you give will be **tripled.**	
[Participant's ID], you chose to give **XX points**	Which became **3*XX points**
[Participant's ID], you now have: **150 - XX points**	[Sham partner's ID], now you have: **points**
	[Sham partner's ID], you can now send points back to [Participant's ID] from what he/she gave you: **3*XX points**. How many points do you want to send back to your partner?
[Participant's ID], your partner sent back **XXX points**, **X%** of the tripled amount that you gave	[Sham partner's ID], you sent back **XXX points**, **X%** of what [Participant's ID] gave to you. You kept **XXX points**, **X%** of it for yourself
[Participant's ID], you now have **XX points**	[Sham partner's ID], now you have **XX points**
[Participant's ID], do you want to deduct points from [Sham partner's ID]? **Every 10 points** you pay deduct **30 points** of [Sham partner's ID]	
[Participant's ID], you chose to pay **XX points** to deduct **XX points** from [Sham partner's ID]	[Sham partner's ID], your partner deducted **XXX points** from you

[Participant's ID], your total is now: **XXX points**	[Sham partner's ID], your total is now: **XXX points**

Congratulations! You have earned **XXX points** as the truster.

[If participants had not played the two roles yet]
Now you and your partner alternate the roles.
You are still paired with Participant [sham partner's ID].

[Counter-balance: If the participants became the responder]

This turn, **you are** <u>**the responder**</u> and your partner is <u>the truster</u>.
Now you and your partner are each given **50 points** (a bonus for you two, regardless of your role).

Responder	*Truster*
[Participant's ID] (you)	**[Sham partner's ID] (your partner)**
50 points	50 points

Responder	*Truster*
[Participant's ID] (you)	**[Sham partner's ID] (your partner)**
50 points	50 points
	[Sham partner's ID], we are giving you **100 points** more. How many points do you want to give to your partner? Whatever you give will be **tripled**.

Responder	*Truster*
[Participant's ID] (you)	**[Sham partner's ID] (your partner)**
50 points	50 points
	[Sham partner's ID], we are giving you **100 points** more.

	How many points do you want to give to your partner? Whatever you give will be **tripled**.
Which became **210 points**	[Sham partner's ID], you chose to give **70 points**

Responder	*Truster*
[Participant's ID] (you)	**[Sham partner's ID] (your partner)**
50 points	50 points
	[Sham partner's ID], we are giving you **100 points** more. How many points do you want to give to your partner? Whatever you give will be **tripled**.
Which became **210 points**	[Sham partner's ID], you chose to give **70 points**
[Participant's ID], you now have: **260 points**	[Sham partner's ID], now you have: **130 points**

Responder	*Truster*
[Participant's ID] (you)	**[Sham partner's ID] (your partner)**
50 points	50 points
	[Sham partner's ID], we are giving you **100 points** more. How many points do you want to give to your partner? Whatever you give will be **tripled**.
Which became **210 points**	[Sham partner's ID], you chose to give **70 points**
[Participant's ID], you now have: **260 points**	[Sham partner's ID], now you have: **130 points**
[Participant's ID], you can now send points back to [Sham partner's ID] from what he/she gave you: XXX points. How many points do you want to send back to your partner?	

How many points do you want to send back to your partner?

- XX points, 0% of what your partner gave you
- XX points, 10% of what your partner gave you

- XX points, 20% of what your partner gave you
- XX points, 30% of what your partner gave you
- XX points, 40% of what your partner gave you
- XX points, 50% of what your partner gave you
- XX points, 60% of what your partner gave you
- XX points, 70% of what your partner gave you
- XX points, 80% of what your partner gave you
- XX points, 90% of what your partner gave you
- XX points, 100% of what your partner gave you

Responder	*Truster*
[Participant's ID] (you)	**[Sham partner's ID] (your partner)**
50 points	50 points
	[Sham partner's ID], we are giving you **100 points** more. How many points do you want to give to your partner? Whatever you give will be **tripled**.
Which became **210 points**	[Sham partner's ID], you chose to give **70 points**
[Participant's ID], you now have: **260 points**	[Sham partner's ID], now you have: **130 points**
[Participant's ID], you can now send points back to [Sham partner's ID] from what he/she gave you: XXX points. How many points do you want to send back to your partner?	
[Participant's ID], you sent back **XXX points**, **X%** of what [Sham partner's ID] gave to you. You kept **XX points**, **X%** of it for yourself	[Sham partner's ID], your partner sent back **XXX points**, **X%** of the tripled amount that you gave

Responder	*Truster*
[Participant's ID] (you)	**[Sham partner's ID] (your partner)**
50 points	50 points
	[Sham partner's ID], we are giving you **100 points** more.

	How many points do you want to give to your partner? Whatever you give will be **tripled**.
Which became **210 points**	[Sham partner's ID], you chose to give **70 points**
[Participant's ID], you now have: **260 points**	[Sham partner's ID], now you have: **130 points**
[Participant's ID], you can now send points back to [Sham partner's ID] from what he/she gave you: XXX points. How many points do you want to send back to your partner?	
[Participant's ID], you sent back **XXX points**, **X%** of what [Sham partner's ID] gave to you. You kept **XX points**, **X%** of it for yourself	[Sham partner's ID], your partner sent back **XXX points**, **X%** of the tripled amount that you gave
[Participant's ID], you now have **XXX points**	[Sham partner's ID], now you have **XXX points**

Responder	*Truster*
[Participant's ID] (you)	**[Sham partner's ID] (your partner)**
50 points	50 points
	[Sham partner's ID], we are giving you **100 points** more. How many points do you want to give to your partner? Whatever you give will be **tripled**.
Which became **210 points**	[Sham partner's ID], you chose to give **70 points**
[Participant's ID], you now have: **260 points**	[Sham partner's ID], now you have: **130 points**
[Participant's ID], you can now send points back to [Sham partner's ID] from what he/she gave you: XXX points. How many points do you want to send back to your partner?	

[Participant's ID], you sent back **XXX points**, **X%** of what [Sham partner's ID] gave to you. You kept **XX points**, **X%** of it for yourself	[Sham partner's ID], your partner sent back **XXX points**, **X%** of the tripled amount that you gave
[Participant's ID], you now have **XXX points**	[Sham partner's ID], now you have **XXX points**
	[Sham partner's ID], do you want to deduct points from [Participant's ID] **Every 10 points** you pay deduct **30 points** of [Participant's ID]

Responder	*Truster*
[Participant's ID] (you)	**[Sham partner's ID] (your partner)**
50 points	50 points
	[Sham partner's ID], we are giving you **100 points** more. How many points do you want to give to your partner? Whatever you give will be **tripled**.
Which became **210 points**	[Sham partner's ID], you chose to give **70 points**
[Participant's ID], you now have: **260 points**	[Sham partner's ID], now you have: **130 points**
[Participant's ID], you can now send points back to [Sham partner's ID] from what he/she gave you: XXX points. How many points do you want to send back to your partner?	
[Participant's ID], you sent back **XXX points**, **X%** of what [Sham partner's ID] gave to you. You kept **XX points**, **X%** of it for yourself	[Sham partner's ID], your partner sent back **XXX points**, **X%** of the tripled amount that you gave
[Participant's ID], you now have **XXX points**	[Sham partner's ID], now you have **XXX points**

	[Sham partner's ID], do you want to deduct points from [Participant's ID] Every 10 points you pay deduct 30 points of [Participant's ID]
[Participant's ID], your partner deducted XXX points from you	[Sham partner's ID], you chose to pay XXX points to deduct XXX points from [Participant's ID]

[If participants return less than 50%, there was a 50% chance that the sham truster pays 20 points to deduct 60 points from the participant]

Responder	*Truster*
[Participant's ID] (you)	[Sham partner's ID] (your partner)
50 points	50 points
	[Sham partner's ID], we are giving you 100 points more. How many points do you want to give to your partner? Whatever you give will be tripled.
Which became 210 points	[Sham partner's ID], you chose to give 70 points
[Participant's ID], you now have: 260 points	[Sham partner's ID], now you have: 130 points
[Participant's ID], you can now send points back to [Sham partner's ID] from what he/she gave you: XXX points. How many points do you want to send back to your partner?	
[Participant's ID], you sent back XXX points, X% of what [Sham partner's ID] gave to you. You kept XX points, X% of it for yourself	[Sham partner's ID], your partner sent back XXX points, X% of the tripled amount that you gave
[Participant's ID], you now have XXX points	[Sham partner's ID], now you have XXX points

	[Sham partner's ID], do you want to deduct points from [Participant's ID] Every 10 points you pay deduct 30 points of [Participant's ID]
[Participant's ID], your partner deducted XXX points from you	[Sham partner's ID], you chose to pay XXX points to deduct XXX points from [Participant's ID]
[Participant's ID], your total is now: XXX points	[Sham partner's ID], your total is now: XXX points

Congratulations! You have earned XXX points as the responder.

[If participants had played the two roles]

This concludes the first block.

Partner switching after the TGP

[High Partner Choice Condition]
Now you are able to switch partners. You can either keep Participant [Sham partner's ID] or switch to a different partner.

Would you like to switch partners?

I would like to:

- Keep the same partner
- Switch to a different partner

[If choosing "Keep the same partner"]
You will continue to interact with your former partner, Participant [Sham partner's ID] in the next block.
[If choosing "Switch to a different partner"]
You now have a different partner, Participant [random two-digit number].

[Low Partner Choice Condition]
You will continue to interact with your former partner, Participant [Sham partner's ID].

After partner switching

The program has decided that there will be no more blocks. This ends this section of your session.

This concludes your interaction with other people.

Appendix B

1. Instructions in study 2a

1-1. Participants' initials for the *Identifiability* manipulation

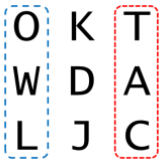
Please enter your initials (the first letters of your first and last names).
We will refer to you by your initials throughout this study session.
For example, if your name is Albus Dumbledore, enter "A.D." Please make sure to include periods (.) after each letter.

1-2. Minimal Group Paradigm

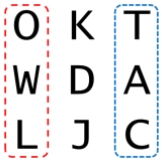
Here's a quick puzzle.

Find a 3-letter word in this picture.
Enter the first 3-letter word you find below.

Some people found OWL first. Other people found CAT first.
Let's call those who found OWL "TEAM Blue" and those who found CAT "TEAM Red".

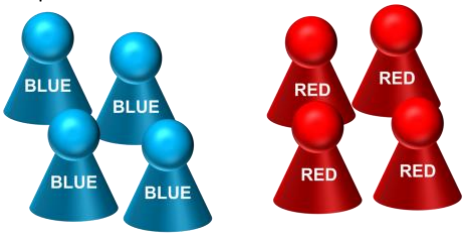


[OR] Let's call those who found CAT "TEAM Blue" and those who found OWL "TEAM Red".



[The combination of the animals and the colors was randomized.]

People on the same team tend to share many personality and cognitive traits.



You are a member of TEAM Blue/Red.

1-3. Introduction to (sham) partners

Hi everybody! Here are all the people participating in this session. You will interact with some of them, but not all of them.

- J.T.
- S.M.
- [Participant's ID]
- D.S.
- K.N.
- A.R.
- N.H.
- M.K.
- S.H.
- D.M.
- M.Y.
- a.w.
- w.z.
- C.F.
- Y.S.
- G.S.
- R.C.

You may want to form impressions of your various partners.
However, in some interactions, you might not know who you are interacting with. You will only know your partner as ?.?.

Sometimes your partner might not know that they are interacting with you. Your partner will only know you as ?.?.

1-4. Instructions before the Dictator Game with Punishment (DGP)

Now you are going to interact with other people who are participating in this study.
Other people will be your partner(s) in various situations in which you can benefit each other.
First, we will give you points that can be used throughout this study.

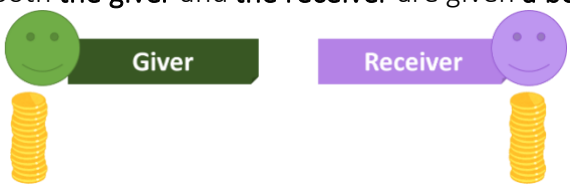
You and your partners can give each other points in various situations.

Imagine that points are something like money – e.g., the program converts points to real money at the end of the study.
The number of points you will earn depends on *both your decisions and your partners' decisions*.
There are no right or wrong decisions. We will explain how everything works before you and your partners make any decisions.
You will be given points at the beginning of each interaction.
During that interaction, you *might lose points, depending on your decisions and your partners' decisions*.
However, since we will give you points for every interaction in which you engage, you will always have a positive number of points ***in total*** by the end of the study.

1-5. Instructions for the DGP

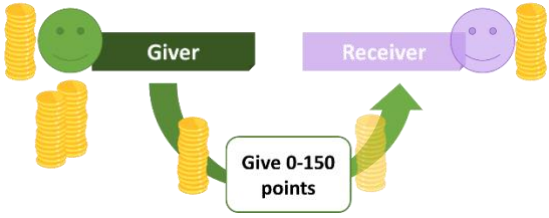
First, you will be interacting with another person, just once.
One of you will be **the giver**, the other will be **the receiver**.

There are **three** steps.
Step 0:
Both **the giver** and **the receiver** are given **a bonus: 50 points**.

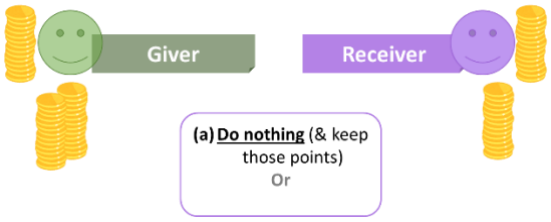


The interaction itself takes **two** steps.
Step 1:
Then **the giver** is given 150 points for this interaction.

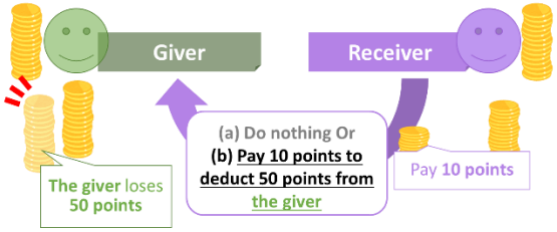
Step 1:
The giver can do anything they want with the 150 points.
The giver can keep all of them (i.e., give 0 points).
Or **the giver** can give points to **the receiver**. **The giver** can give any number of points to **the receiver**—none, some, all—from 0-150 in 10-point increments.



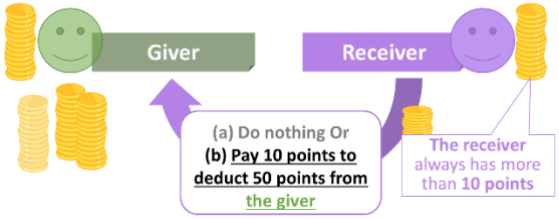
Step 2:
After receiving points from **the giver**, **the receiver** can either:
(a) do nothing (and keep the points they received), **OR**



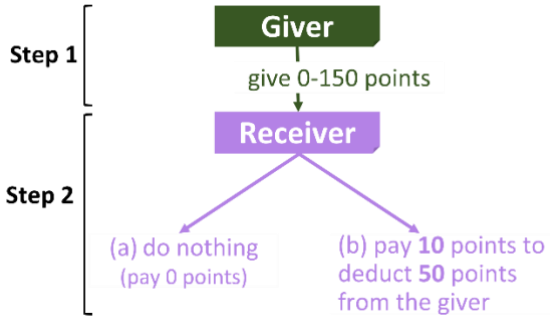
Step 2:
After receiving points from the **giver**, **the receiver** can either:
(a) do nothing (and keep the points they received), **OR**
(b) pay **10 points** to deduct **50 points** from **the giver**.



Remember that both **the giver** and **the receiver** first received **50 points** in Step 0.
Therefore, **the receiver** can always afford to pay 10 points (if **the receiver** wants to).



Here is a summary of the interaction.



After you have interacted once, **with one person**, you will interact again with **another, different person**. You will interact with **several different partners**.
Your **role** may change when you interact with **the new partner**.

Would you like to read the explanation again?

- Yes

- No

[If Yes, go back to the first page of the DGP instruction]

1-6. Two practice rounds of the DGP

Let's practice before you actually interact with your partner.
You will practice several times, but note that you do not actually win or lose points during the practice session.
After the practice session, you will be given actual points and will be able to earn more points with it.

You are now paired with your practice partner.
You are the giver and your practice partner is the receiver.
…
[See "Instruction during the DGP" below for details. Participants experienced two practice rounds of the DGP, once as the giver, once as the receiver. They always practiced the giver first. The sham practice receiver never deducted points from the participant as long as the participants had given them 60 points or more, but if the participant had given them 50 points or less the sham practice receiver always paid 10 points to deduct 50 points from the participant. The sham practice giver either (i) kept 140 points and gave the participants 10 points or (ii) kept 70 points and gave the participant 80 points.**]**

This concludes the practice session.

1-7. Comprehension check questions for the DGP

Here are a few quizzes to help you understand.
Please read the questions carefully and choose the best answer.

Q1. In the first step, the giver divides 150 points between themself and the receiver.
When the giver gets 150 points, how many points must the giver give to the receiver?

- The giver must give half the points
- The giver must keep all the points
- The giver can give any number, from 0 to 150 points

[If choosing the correct answer, participants proceeded to the next question. Otherwise, participants were told their answer was not correct and then given a chance to answer the same question again.]
Your answer is correct! The giver can give any number of points to the receiver—none, some, all—from 0-150.
[If missing the same question twice]
Your answer is not correct. The giver can give any number of points to the receiver—none, some, all—from 0-150. So the correct answer is "The giver can give any number, from 0 to 150 points".

Q2. If the giver gives 50 of the 150 points, how much does the giver keep?

- The giver keeps **50 points** for themself
- The giver keeps **150 points** for themself
- The giver keeps **100 points** for themself

[If choosing the correct answer]
Your answer is correct!
If the **giver** gives 50 points to **the receiver**, **the giver** keeps the rest of 150 points: **100 points**.
[If missing the same question twice]
Your answer is **not correct**. If the **giver** gives 50 points to **the receiver**, **the giver** keeps the rest of 150 points: **100 points**. So the correct answer is "The giver keeps **100 points** for themself".

Q3. In the second step, after receiving the points from the giver, the receiver can either (a) do nothing (pay 0 points), or (b) pay 10 points.
What happens to the giver if the receiver pays 10 points?

- The giver loses **50 points** (but the receiver does not get these 50 points)
- The giver loses **50 points** and then **the receiver earns 50 points**
- Nothing happens. Each keeps the points they have.

[If choosing the correct answer]
Your answer is correct! If the receiver **pays 10 points**, **the giver loses 50 points**. (But the receiver does not get these 50 points.)
[If missing the same question twice]
Your answer is **not correct**. If the receiver pays 10 points, the giver loses 50 points. (But the receiver does not get these 50 points.) So the correct answer is "The giver loses **50 points** (but the receiver does not get these 50 points)".

Q4. What happens to the giver if the receiver does nothing?

- The giver loses **50 points**
- The giver loses **30 points**
- Nothing happens. Each keeps the points they have.

[If choosing the correct answer]
Your answer is correct! If the receiver does nothing and pays 0 points, the giver does not lose points (nothing happens).
[If missing the same question twice]
Your answer is **not correct**. If the receiver **does nothing and pays 0 points**, **the giver does not lose points (nothing happens)**. So the correct answer is "Nothing happens. Each keeps the points they have."

1-8. Instruction before the DGP with partners

Now you are going to **actually interact with your partner**.

Again, after you have interacted once, with one person, you will interact with another, different person. You will interact with **a number of different partners**.

Remember that you are a member of TEAM Red/Blue.

[OR]

Looking for a partner...
[4 seconds of wait time]
You are now paired with your partner.

1-9. Dummy rounds of DGP with unknown partners (without Anonymity IV)

[Participants played 10 rounds of DGP, each with a different sham partner. In round 1 and 6, participants interacted with an "unknown" generous sham giver who was anonymous to the participant; the sham partner "knew" the participant's identity. These givers were always generous and either (i) kept either 80 points and gave 70 points to the participant or (ii) kept 70 points and gave 80 points to the participant. One of the unknown generous partners was ingroup, the other was outgroup (random order).

Round	Group membership of partner	Participant's identity to the partner	Partner's role
1	Either Ingroup or Outgroup	Identified, partner is anonymous	Giver
6	Either Ingroup or Outgroup	Identified, partner is anonymous	Giver

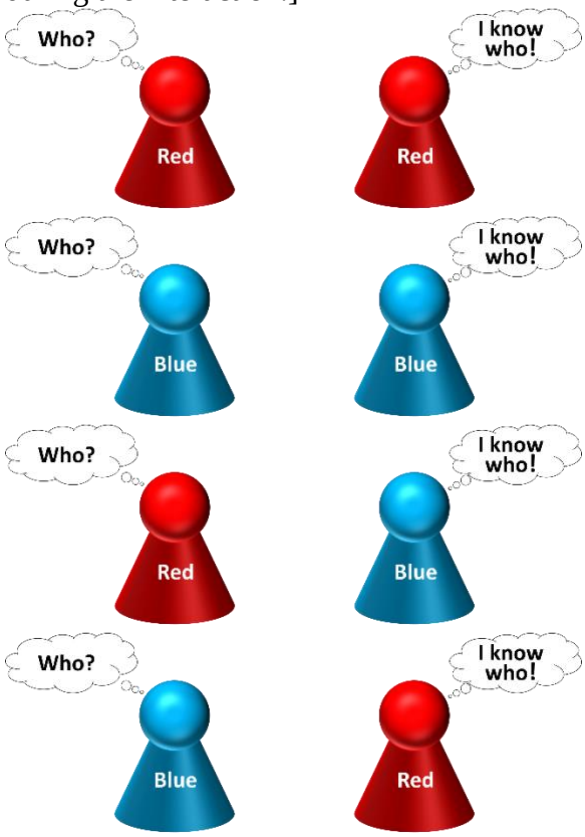
]

[A round with an unknown generous sham giver]
[Participant's initial] and ?.?., now you two are paired.
[Participant's initial] is on TEAM Red/Blue.
?.?. is on TEAM Red/Blue.

[Participant's initials], you do NOT know who your partner is.
That's why you do not see your partner's initials. You only know that ?.?. is on TEAM Red/Blue.

?.?., you know who your partner is: Your partner is [Participant's initial]. You also know that [Participant's initial] is on TEAM Red/Blue.

[One of the below four figures was shown. The initials of the participant and the sham partner were shown just beneath the corresponding silhouette (the anonymous partner's initials were shown as "**?.?.**") The same figure and the initials appeared on every page during the interaction.]



Your partner, **?.?.**, was given 150 points.
Your partner, **?.?.**, decided to **keep** 80/70 **points** and **offer you** 70/80 **points**.

[PUNISHMENT DECISION: with a generous unknown giver (non-DV)]
Your partner, **?.?.**, kept 80/70 points. You received 70/80 points.
What would you like to do?

- Pay 10 points to deduct 50 points from your partner
- Do nothing (pay 0 points)

[Feedback if the participants decided to punish]
Your partner, **?.?.**, kept 80/70 points. You received 70/80 points.
You paid 10 points to deduct 50 points from your partner.
Therefore, your partner earned 30/20 points and **you earned** 60/70 **points**.

[Feedback if the participants decided not to punish]

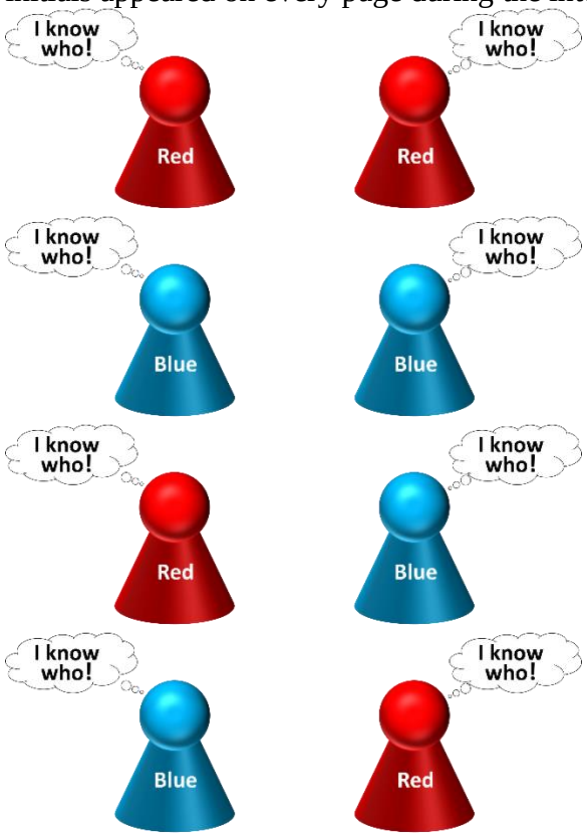
Your partner, ?.?., kept 80/70 points. You received 70/80 points.
You did not deduct points from your partner.
Therefore, your partner earned 80/70 points and **you earned 70/80 points.**

1-10. Eight rounds of DGP (with Anonymity IV manipulation)

[Participants played 10 rounds of DGP, each with a different sham partner. In eight rounds except for round 1 and 6, that is, in round 2, 3, 4, 5, 7, 8, 9, and 10, participants interacted with one of the eight partners below in a random order:

[When participant was IDENTIFEID by the partner]
[Participant's initial] and [Sham partner's initial], now you two are paired.
[Participant's initial] is on TEAM Red/Blue
[Sham partner's initial] is on TEAM Red/Blue.

[One of the below four figures was shown. The initials of the participant and the sham partner were shown just beneath the corresponding silhouettes. The same figure and the initials appeared on every page during the interaction.]

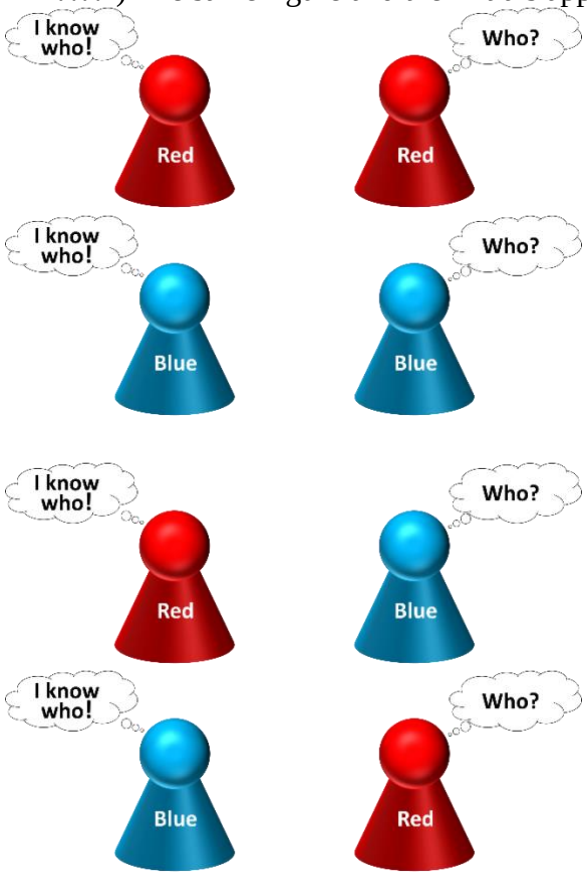


[OR: When participant was ANONYMOUS to the partner]
?.?. and [Sham partner's initial], now you two are paired.
?.?. is on TEAM Red/Blue
[Sham partner's initial] is on TEAM Red/Blue.

?.?., you know who your partner is: Your partner is [Sham partner's initial]. You also know that [Sham partner's initial] is on TEAM Red/Blue.

[Sham partner's initial], you do NOT know who your partner is.
That's why you do not see your partner's initials. You only know that ?.?. is on TEAM Red/Blue.

[One of the below four figure was shown. The initials of the participant and the sham partner were shown just beneath the corresponding silhouette (the participant's initials were shown as "**?.?.**") The same figure and the initials appeared on every page during the interaction.]



[When the participant was the GIVER:
The sham receiver conditionally punished the participant. If the participant gave the sham receiver 60 points or more, the sham receiver did not pay 10 points to deduct 50 points from

the participant. If the participant gave the sham receiver 50 points or less, there was a 50% chance that the sham receiver pays 10 points to deduct 50 points from the participant. Below, when either of them was anonymous, the initials were shown as "**?.?.**"]

[Participant's initial], you are the giver.

[COOPERATION DV]
You are given 150 points.
How would you like to divide it?
(For you, for your partner: [Sham partner's initial])

- (150 points, 0 points)
- (140 points, 10 points)
- (130 points, 20 points)
- (120 points, 30 points)
- (110 points, 40 points)
- (100 points, 50 points)
- (90 points, 60 points)
- (80 points, 70 points)
- (70 points, 80 points)
- (60 points, 90 points)
- (50 points, 100 points)
- (40 points, 110 points)
- (30 points, 120 points)
- (20 points, 130 points)
- (10 points, 140 points)
- (0 points, 150 points)

Waiting for your partner...
[3-5 seconds of wait time]

[Feedback if the participant was punished (50% of the time when participants gave the partner 50 points or less)]
This concludes your interaction with the current partner.
You decided to keep [X] points and offer your partner [150-X] points.
Your partner paid 10 points to deduct 50 points from you.
Therefore, you earned [X-50] points and your partner earned [150-X-10] points.

[Feedback if the participant was not punished (when participants gave the partner 60 points or more)]
This concludes your interaction with the current partner.
You decided to keep [X] points and offer your partner [150-X] points.

Your partner did not deduct points from you.
Therefore, you earned [X] points and your partner earned [150-X] points.

[When the participant was the RECEIVER:
The sham givers were always stingy. They either (i) kept 100 points and gave the participant 50 points or (ii) kept 120 points and gave the participant 30 points.
Below, when either of them was anonymous, the initials were shown as "**?.?.**"]

[Participant's initial], you are the receiver.

Waiting for your partner...
[3-5 seconds of wait time]
Your partner, [Sham partner's initial], was given 150 points.
Your partner, [Sham partner's initial], decided to keep 100/120 points and offer you 50/30 points.

[PUNISHMENT DV]
Your partner, [Sham partner's initial], kept 100/120 points. You received 50/30 points.
What would you like to do?

- Pay 10 points to deduct 50 points from your partner
- Do nothing (pay 0 points)

[Feedback if the participants decided to punish]
This concludes your interaction with the current partner.
Your partner, [Sham partner's initial], kept 100/120 points. You received 50/30 points.
You paid 10 points to deduct 50 points from your partner.
Therefore, your partner earned 50/70 points and you earned 40/20 points.

[Feedback if the participants decided not to punish]
This concludes your interaction with the current partner.
Your partner, [Sham partner's initial], kept 100/120 points. You received 50/30 points.
You did not deduct points from your partner.
Therefore, your partner earned 100/120 points and you earned 50/30 points.

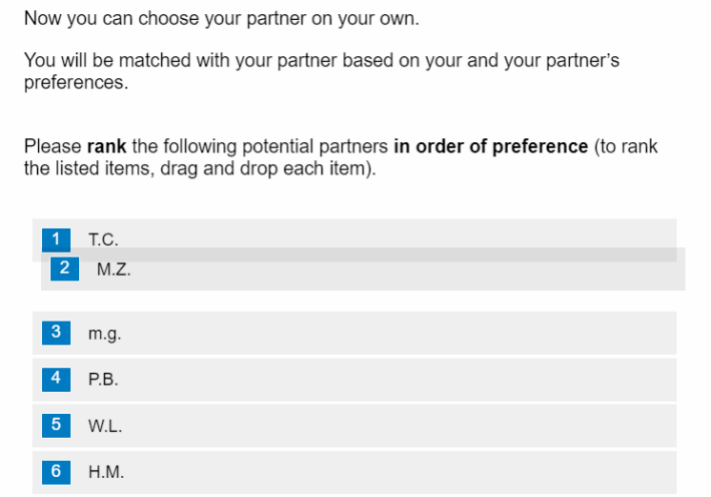
[Between every round]
Looking for a new partner...
[3-7 seconds of wait time]
You are now paired with a new partner.

[After 10 interactions]
This concludes your interaction with other people.
Congratulations! You have earned XXX points in total.

2. Instructions unique to studies 2b and 2c
2-1. Introduction to (sham) partners before the DGP
[In studies 2b and 2c, participants were introduced to sham partners just before the DGP interactions, not before the instructions for DGP.]
Hi everybody! Here are all the people participating in this session. You will interact with some of them, but not necessarily all of them.
Note that some of you may be in different places.
[The number of sham partners was reduced to 8, a realistic number for study 2c, which was conducted in a university lab under social distancing guideline for COVID-19 outbreak.]

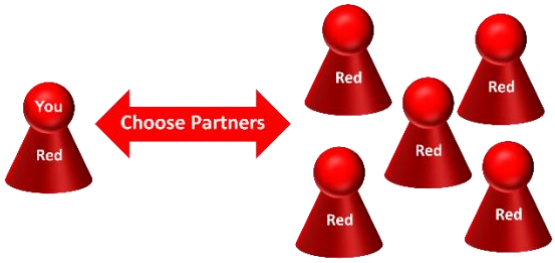
- J.T.
- S.H.
- N.H.
- M.K.
- G.S.
- Y.S.
- [Participant's ID]
- A.R.
- S.H.

Like in the practice, you will be randomly paired with your partners in the first several interactions.
But later, you will be able to choose who you would like to interact with (like shown below).

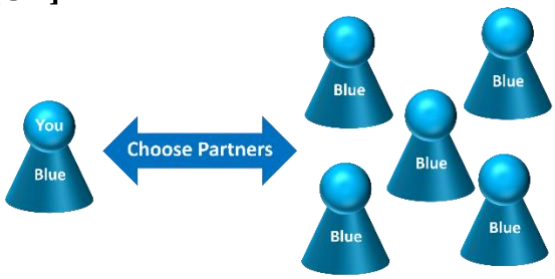


[*Ingroup Partner Choice* condition]
People on TEAM Red/Blue will choose their partners from TEAM Red/Blue; people on TEAM Blue/Red will choose their partners from TEAM Blue/Red.

So you will choose your partner from TEAM Red/Blue.
At the same time, those who are on TEAM Red/Blue will choose you as a partner.



[OR]

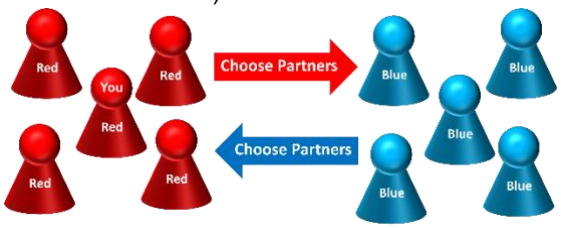


[OR: ***Outgroup Partner Choice*** **condition**]

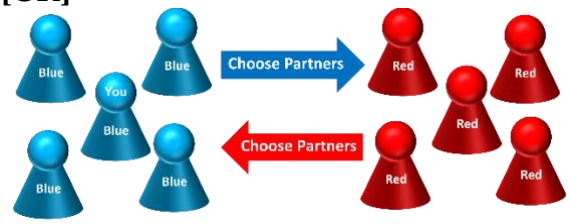
People on TEAM Red/Blue will choose their partners from TEAM Blue/Red; people on TEAM Blue/Red will choose their partners from TEAM Red/Blue.

So you will choose your partner from TEAM Blue/Red.

At the same time, those who are on TEAM Blue/Red will choose you as a partner.



[OR]



[OR, only in study 2c: ***Universal Partner Choice condition***]

[In the Universal partner choice condition, there was no additional interaction about partner choice mentioning groups.]

[The anonymity IV instruction similar to the one in study 2a]

However, in some interactions, you might not know who you are interacting with. You will only know your partner as ?.?.

Sometimes your partner might not know that they are interacting with you. Your partner will only know you as ?.?.

So when you know who you are interacting with, you may want to pay attention to their names.

2-2. Ranking partners

[*Ingroup Partner Choice* condition]

Now you can choose your partner from TEAM Red/Blue.
Here are people on TEAM Red/Blue who you have interacted with.

- [Ingroup partner initials 1]
- [Ingroup partner initials 2]
- [Ingroup partner initials 3]
- [Ingroup partner initials 4]

You will be matched with your partner based on your and your partner's preferences.
Who would you like to interact with? Please rank the following potential partners in order of preference (to rank the listed items, drag and drop each item).

- [Ingroup partner initials 1]
- [Ingroup partner initials 2]
- [Ingroup partner initials 3]
- [Ingroup partner initials 4]

[OR: ***Outgroup Partner Choice*** **condition**]
Now you can choose your partner from TEAM Blue/Red.
Here are people on TEAM Blue/Red who you have interacted with.

- [Outgroup partner initials 1]
- [Outgroup partner initials 2]
- [Outgroup partner initials 3]
- [Outgroup partner initials 4]

You will be matched with your partner based on your and your partner's preferences.
Who would you like to interact with? Please rank the following potential partners in order of preference (to rank the listed items, drag and drop each item).

- [Outgroup partner initials 1]
- [Outgroup partner initials 2]
- [Outgroup partner initials 3]
- [Outgroup partner initials 4]

[OR: ***Universal Partner Choice condition***]
Now you can choose your partner.
Here are people who you have interacted with.

- [Ingroup partner initials 1]
- [Ingroup partner initials 2]
- [Ingroup partner initials 3]
- [Ingroup partner initials 4]
- [Outgroup partner initials 1]
- [Outgroup partner initials 2]
- [Outgroup partner initials 3]
- [Outgroup partner initials 4]

[The order of ingroup and outgroup partners were randomized.]

You will be matched with your partner based on your and your partner's preferences.
Who would you like to interact with? Please rank the following potential partners in order of preference (to rank the listed items, drag and drop each item).

- [Ingroup partner initials 1]
- [Ingroup partner initials 2]
- [Ingroup partner initials 3]
- [Ingroup partner initials 4]
- [Outgroup partner initials 1]
- [Outgroup partner initials 2]
- [Outgroup partner initials 3]
- [Outgroup partner initials 4]

[The order of ingroup and outgroup partners were randomized.]

2-3. One DGP round with a chosen partner

[Participant's initial] and [The top choice sham partner's initial], now you two are paired.
[In the interaction with the top choice partner, neither the partner nor the participant was anonymous. The top choice partner always played the giver and generously gave participants 80 points out of 150.]

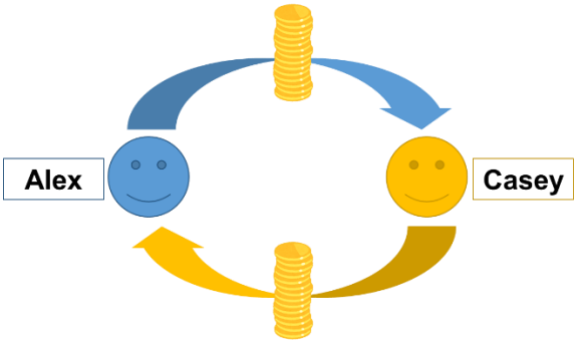
[After one round with the top choice partner]

This concludes your interaction with other people.

Appendix C

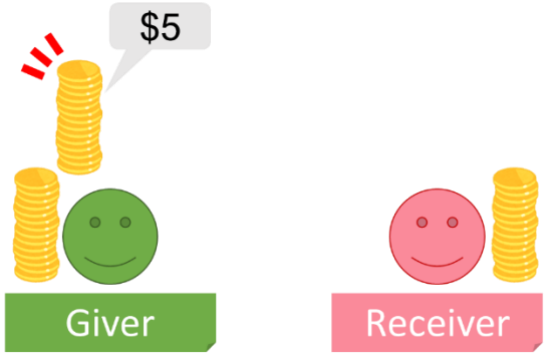
Study 3a: Instructions
1. Dictator Game with Taking Option

[Page 1]
You will observe two people, Alex and Casey, interact with each other.
Imagine Alex and Casey are participating in a study like you are.
They will have an opportunity to benefit each other.



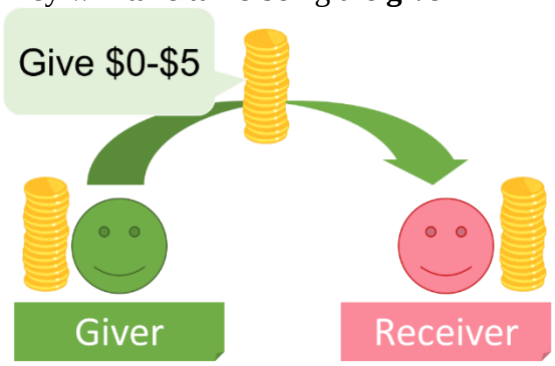
[Page 2]
Before you observe **Alex** and **Casey** interacting, here's how two people can benefit each other.
In the beginning of the interaction, two of them are each given $5.

[Page 3]
One of them becomes a **giver**, the other becomes a **receiver**.
Then the **giver** receives an additional $5.

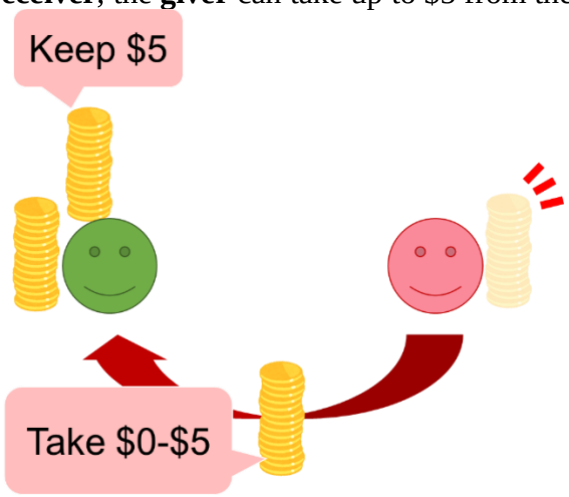


[Page 4]

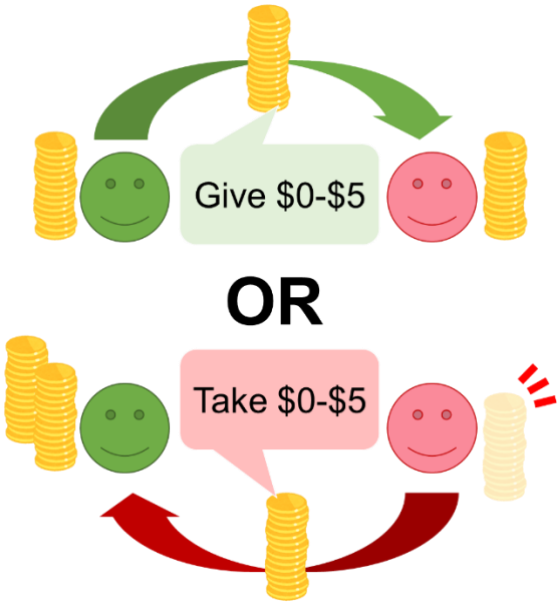
The **giver** can share the additional $5 with the **receiver**. That is, the **giver** can give the **receiver** none, some, or all of the $5—as much as they would like to.
They will take turns being the **giver**. That's why they can benefit each other.



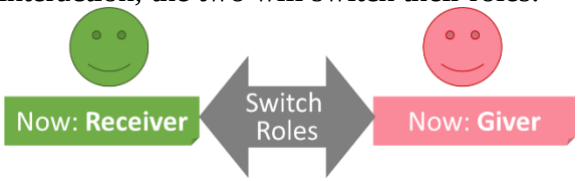
[Page 5]
However, the **giver** can also take money from the **receiver**. Instead of sharing $5 with the **receiver**, the **giver** can take up to $5 from the **receiver**.



[Page 6]
To sum up, the **giver** can either give money to the **receiver**—from $0 to $5—or take money from the **receiver**—from $0 to $5.

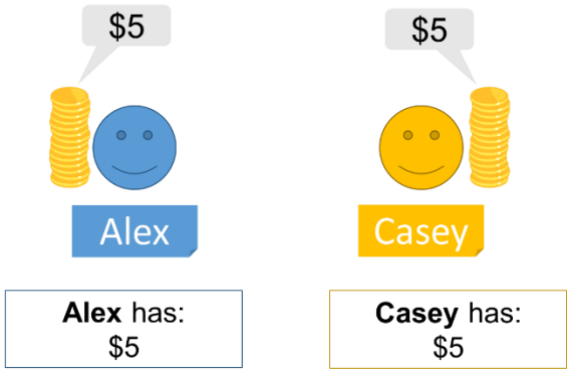


[Page 7]
The two repeat the interaction, taking turns being the **giver** and the **receiver**. In every new interaction, the two will switch their roles.

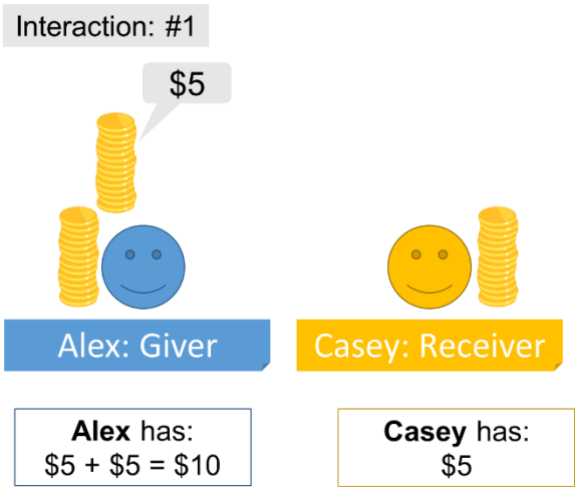


[Page 8]
Now you are going to see how **Alex** and **Casey** treated each other. They know that they will interact repeatedly.
This is their 1st interaction.
First, **Alex** and **Casey** were given $5 each.

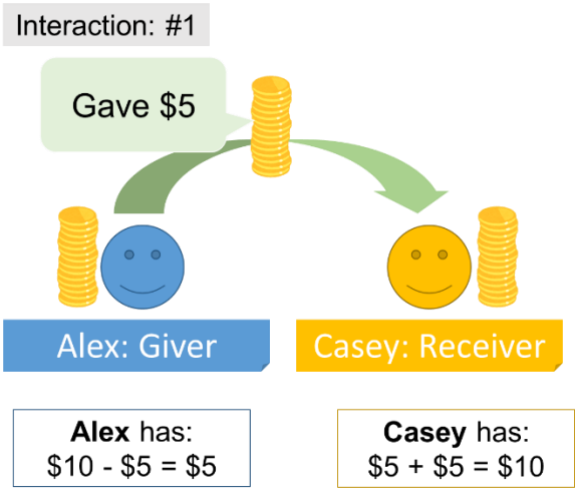
Interaction: #1



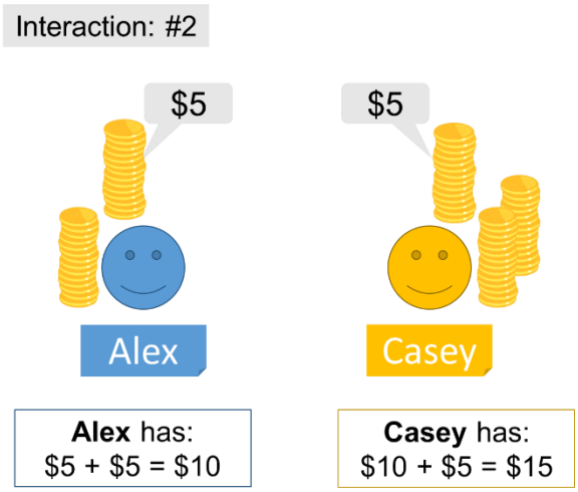
[Page 9]
This time, **Alex** is the **giver** and **Casey** is the **receiver**.
So **Alex** received an additional $5.



[Page 10]
Alex decided to give **Casey** $5.
That's how their 1st interaction ended.

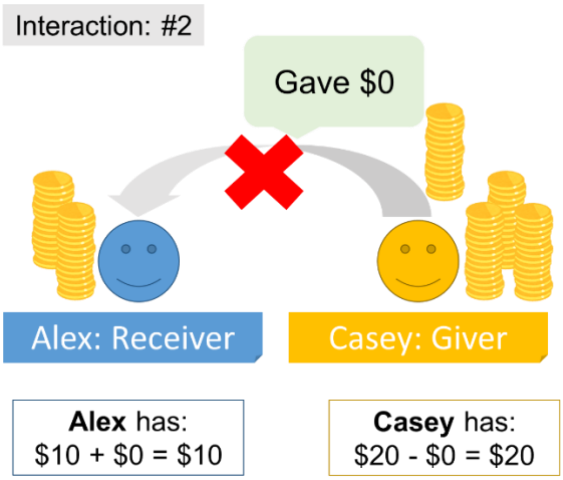


[Page 11]
This is their 2nd interaction.
As before, **Alex** and **Casey** were given $5 each to start with.

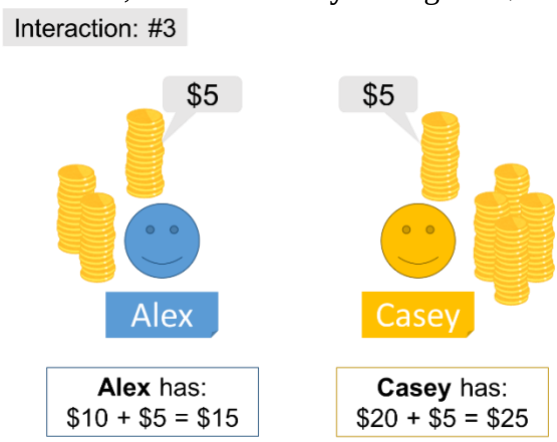


[Page 12]
This time, **Casey** is the **giver** and **Alex** is the **receiver**.
So **Casey** received an additional $5.

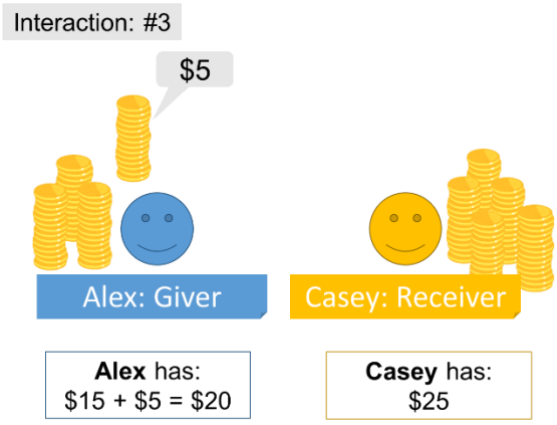
[Page 13]
Casey decided to give **Alex** $0.
That's how their 2nd interaction ended.



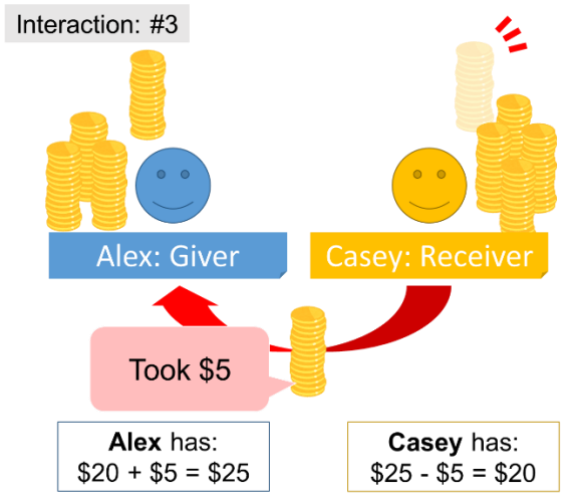
[Page 14]
This is their 3rd interaction.
As before, **Alex** and **Casey** were given $5 each to start with.



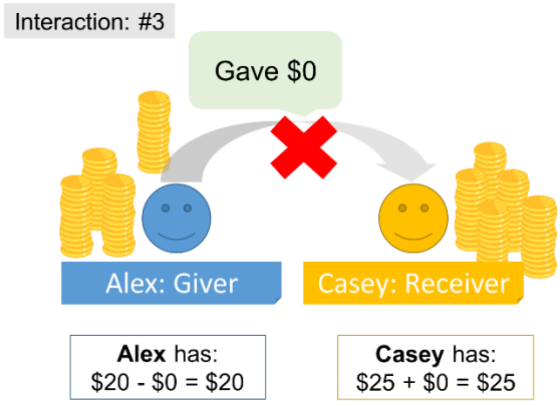
[Page 15]
This time, **Alex** is the **giver** again and **Casey** is the **receiver**.
So **Alex** received an additional $5.



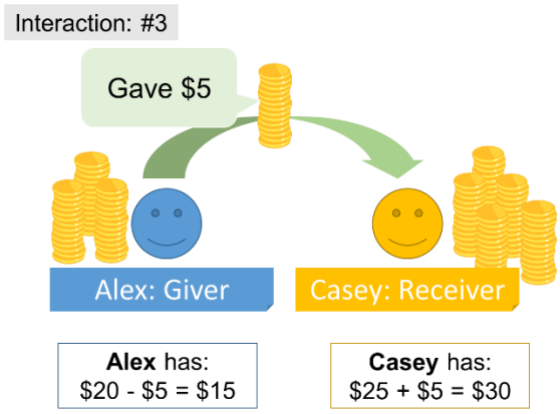
[Page 16. **Each participant saw only one of the three panels below.]**
[1. Punish condition]
Alex decided to take $5 from **Casey**.
That's how their 3rd interaction ended.



[2. Withdraw condition]
Alex decided to give **Casey** $0.
That's how their 3rd interaction ended



[3. No negative sanction condition]
Alex decided to give **Casey** $5.
That's how their 3rd interaction ended.



2. Attention check questions
[Participants who could now correctly respond to (i) the third question or/and (ii) both the first and the second questions were excluded from the study.]

Here are a few quiz questions about what has happened.
Please read the questions carefully and choose the best answer.

1. In their first interaction, how much money did **Alex** give **Casey**?
- Gave $0
- Gave $1
- Gave $2
- Gave $3
- Gave $4
- Gave $5

2. When it was **Casey's** turn to be the giver, how much money did **Casey** give to **Alex**?
- Gave $0
- Gave $1
- Gave $2
- Gave $3
- Gave $4
- Gave $5

3. How did **Alex** respond when **Casey** gave **Alex** nothing?
- **Alex** gave **Casey** $5
- **Alex** gave **Casey** $0
- **Alex** took $5 from **Casey**

3. Evaluating reputations
Given what you saw, tell us what you think about **Alex**.
How **exploitable** do you think **Alex** is?
- 1. Not at all

- 2.
- 3.
- 4. Average
- 5.
- 6.
- 7. Extremely

[The same question for the rest of the adjectives (the question order was randomized): weak, gullible, unwise, incompetent, vengeful, aggressive, impulsive, cowardly, frightened, mean, and careless, dependable, likable, forgiving, generous, considerate, cooperative, trustworthy, honorable, friendly, kind, fair, and emotionally-stable.]

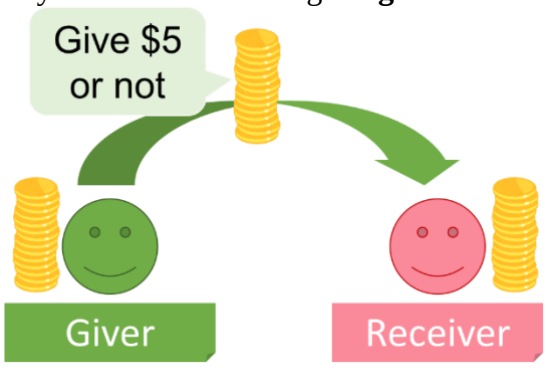
4. Partner choice preference
Imagine you have to engage in the same kind of interaction.
How much would you like to interact with **Alex**?

- 1. Not at all
- 2. Slightly
- 3. Somewhat
- 4. Moderately
- 5. Extremely

2. Study 3b: New instructions to participants

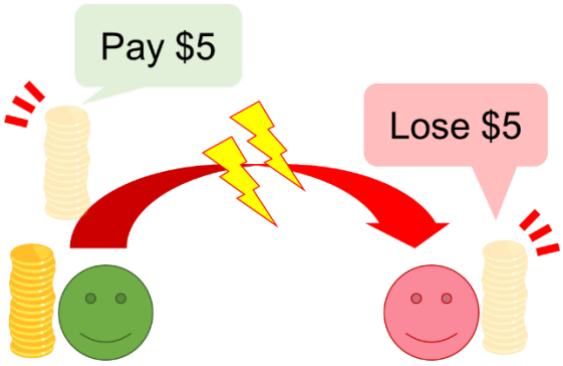
[In study 3b, the instruction for the interaction was identical with study 3b, except for page 4, 5, 6 (instructions), and 16 in Punish condition (Alex pay a cost to punish Casey). Red text indicates the changes from study 3a.']

[Page 4]
The **giver** can share the additional $5 with the **receiver**. That is, the **giver** can decide whether to give the receiver $5 or not.
They will take turns being the **giver**. That's why they can benefit each other.



[Page 5]

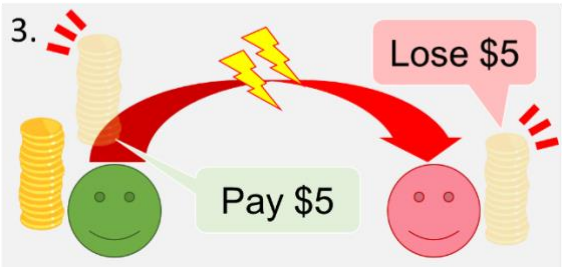
However, the **giver** can also reduce the **receiver**'s earnings. Instead of sharing $5 with the **receiver**, the **giver** can pay $5 to reduce the receiver's earnings by $5.



[Page 6]

To sum up, the **giver** has 3 options:

1. gives the **receiver** $5
2. gives the **receiver** nothing ($0)
3. pays $5 to reduce the **receiver**'s earnings by $5.



[Page 16. **Costly punish condition]**
Alex decided to pay $5 to reduce **Casey**'s earnings by $5.
That's how their 3rd interaction ended.

Interaction: #3