

Rewards of Risky Interdependence in Same- and Cross-Race Interactions:
Inducing Trust via High-Stakes Cooperation

by

Connery Knox

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Master of Arts

in

Psychology

Waterloo, Ontario, Canada, 2025

© Connery Knox 2025

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners. I understand that my thesis may be made electronically available to the public.

Statement of Contributions

I joined this project after the study was designed, and data was collected. Data was analyzed by me. All drafts of this thesis were primarily written by me, with Hilary Bergsieker providing edits and input. Aneek Romana assisted with table formatting and references.

Abstract

Mutual trust is vital for interpersonal relationships but often hard to build. Five experiments ($N = 1001$) grounded in Interdependence Theory tested a novel method to boost trust—beyond mere liking—between strangers. In Studies 1-3, Canadian undergraduates completed closeness-building tasks in same- and cross-race dyads, followed by an either “risky” (real money, visible choices) or “safe” (no money, secret choices) iterative prisoner’s dilemma game. The risky game consistently elicited more cooperation and increased trust (twice as much as liking). These trust gains (a) emerged for both same- and cross-race dyads, (b) were mediated by increased cooperation, and (c) generalized to a subsequent negotiation task, where the riskier scenario reduced exploitation concerns and behavioural awkwardness. In Studies 4 and 5, participants forecasting this exact procedure underestimated how much risky cooperative tasks build trust. Together, these findings demonstrate benefits of high-stakes interdependence for establishing interpersonal trust, even across racial group boundaries.

Acknowledgements

Thanks to everyone who contributed to this research, supported me during this adventure, and taught me everything I know about research. To be more specific, thanks Hilary Bergsieker for your patience, support, and mentoring throughout my MA, and your help improving this document through so many iterations. Thanks Megan McCarthy and John Rempel, for your willingness to read through my thesis and offer feedback. Thanks also to Richard Eibach, Denise Marigold, and the graduate students in SPI lab for their feedback on an earlier version of this manuscript. It's been an honor to work with the incredible graduate students of the social area; you have all impacted my work, sometimes through offhand comments and conversations, and sometimes through helping me figure out R code that I should really know by now. Thanks also to Aneek Romana for her work on essentially every time-consuming task that I did not want to do in this project.

I feel an incredible amount of gratitude to have had my partner, Rachel Leong, beside me offering her unwavering support throughout my MA. Whenever my work tired me out, you were always there to provide a seemingly endless amount of understanding, warmth, and love. I appreciate everything you do for me.

I also want to thank my friends outside of the psychology department for listening to me talk about statistics, tables, and research in general. I am very lucky that the most caring, funny, and smart people I have ever met became my friends years ago and stuck around this long. Thanks also to my family for their support and understanding, especially for the last few months leading up to submitting this thesis.

Finally, I want to thank Takuya Nakamura for his incredible music sets on The Lot Radio that I essentially listened to on repeat during the last couple of weeks of writing.

Table of Contents

Author’s Declaration..... *ii*

Statement of Contributions..... *iii*

Abstract.....*iv*

Acknowledgements..... *v*

List of Figures.....*ix*

List of Tables..... *1*

CHAPTER 1: REWARDS OF RISKY INTERDEPENDENCE..... **2**

 Building Trust..... 3

 Using a Social Dilemma to Build Trust..... 4

 Overview of Experiments 6

 Transparency and Openness..... 8

CHAPTER 2: METHOD..... **9**

 Participants and Recruitment..... 9

 Procedure..... 11

 Materials..... 14

 Analytic Approach..... 22

CHAPTER 3: RESULTS..... **25**

 Stakes Effects on Experiencers..... 25

 How Did Forecasters’ Predictions Compare to Experiencers’ Reports?..... 29

Specificity: Did Game Stakes Affect Trust More Than Liking?	32
Extending Gains to a Novel Situation: Negotiation Effects (Study 2-3)	35
Does Cooperation (Actual or Predicted) Mediate Effects of Stakes on Trust?	37
CHAPTER 4: DISCUSSION	47
Persisting Trust Gained Following Cooperation	47
Forecasters Underestimate Cooperation Rates	49
Specific Gains on Trust Versus Liking	51
Similar Effects for Same- and Cross-race Dyads	52
Information or Payment: Which Factor Drives These Effects?	53
Conclusions	54
References	55
Appendix A: Pilot Study Results	66
Appendix B: Additional Materials and Measures	68
Appendix C: Scale Items and Reliabilities	70
Appendix D: Analysis of Moderation by PD Game Matrix	73
Appendix E: Analysis of Secondary Variables	74
Appendix F: Behavioural Coding of Study 3 Experiencers	75
Appendix G: Moderation Analyses	77
Appendix H: Within-participant analysis of Study 4	82
Appendix I: Study-Specific Effects on Key Variables	83

Appendix J: Nesting Structure of the Data..... 85

Appendix K: Correlational Analyses for Key Variables 86

Appendix L: Tests of Closeness Moderation 88

List of Figures

Figure 1: Forecasted Versus Experienced Stakes Effects on Trust Outcomes	30
Figure 2: Post-game Trust and Liking for Experiencers and Forecasters	32
Figure 3: Stakes Effect on Negotiation Trust Mediated by PD Game Trust	36
Figure 4: Experiencers' and Forecasters' Post-Game Trust by Partner's Full Cooperation	38
Figure 5: Theoretical Stakes by Cooperation by Role Mediation on Trust	39
Figure 6: Cooperation by Stakes on Post-game Trust, Split by Role	41
Figure 7: Stakes Predicting Trust Mediated by Partner Cooperation for Experiencers	43
Figure 8: Stakes Predicting Trust Mediated by Partner Cooperation for Forecasters	44
Figure 9: Cooperation by Role on Post-game Trust in the Risky Condition	45

List of Tables

Table 1: Study features	8
Table 2: Exclusions by Study	9
Table 3: Participant Demographic Information by Study	10
Table 4: Binary Cooperation Measures by Stakes and Role	25
Table 5: Continuous Cooperation Measures by Stakes and Role	26
Table 6: Stakes Effects on Trust and Relationship Outcomes	28
Table 7: Stakes Effects on Experiencers' Post-Game (vs. Pre-Game) Outcomes	33
Table 8: Main Effects of Stakes on Negotiation Outcomes	35
Table 9: Mediation Path Coefficients for Trust and Relationship Outcomes	42

CHAPTER 1: REWARDS OF RISKY INTERDEPENDENCE

Imagine leading a diverse organization—be it a business, school, or team—and striving to strengthen relationships between members from different backgrounds. Which strategy would achieve the deepest bonds between members: a potluck lunch, a golf tournament, or a ropes course? All three approaches to increasing cohesion have intuitive appeal, but their psychological underpinnings diverge, respectively relying on cooperation, competition, and high-stakes interdependence (i.e., relying on someone to help you when they have the potential to hurt you, cf. strain-test situations in close relationships; Holmes, 1981). Only on the ropes course, suspended in air by a partner's belay, do people learn to take risks with the expectation that they can count on others to look after them.

Trust, defined as the willingness to accept vulnerability based on confident expectations of an actor's positive intentions or behaviour toward oneself (Rousseau et al., 1998), is crucial for enabling societal coordination (Hardin, 2002), yet creating it—especially within diverse societies—remains a persistent problem (Putnam, 2007). Sustaining a cohesive, coordinated society requires trust among its members: Trust enables individuals to routinely rely upon each other and make personal sacrifices such as paying taxes and obeying laws for collective welfare (Hardin, 2002; Hart, 1988). Trust in fellow citizens is a key determinant of social capital (Bourdieu, 1983) and is linked to improved economic performance (Putnam, 1993), reduced violent crime (Sampson et al., 1997), and greater civic engagement (Putnam, 2000).

Trust matters most when people have divergent motives (Balliet & Van Lange, 2013; Holmes, 2002), conditions that arise frequently in pluralistic societies, making cohesion difficult to achieve. For example, trust levels between individuals from the same group tend to be high (e.g., Brewer, 2001; Insko & Schopler, 1998) relative to trust across group boundaries which

tends to be lower (e.g., Dovidio et al., 2002; Foddy et al., 2009; Vorauer, 2006), and outgroup trust is particularly low among ethnic minorities in Canada (Phan, 2008). The current political climate in Canada and the USA has also created a great deal of uncertainty and, accompanied by this uncertainty, a lack of trust in institutions to act in the populations' best interest (Korzinski, 2022). Additionally, Canadians' levels of trust in each other has declined in the last decade. In 2013 54% of Canadians reported that "most people can be trusted," however in 2024 this has fallen to 44% of Canadians (Government of Canada, 2015, 2024). Overall, Canadians are in a position where being able to build up trust in one another (even across group boundaries) has the potential to be especially impactful on their well-being and belonging, as trusting your neighbors was found to be associated with higher life satisfaction and belonging (Government of Canada, 2023).

At an individual level, establishing trust facilitates including a partner or friend in one's self-concept, which can help people reap the rewards of closer, more meaningful relationships (Drigotas et al., 1999). Expecting favourable treatment from a partner activates the goal to connect (Murray & Holmes, 2009). The Self-Expansion Model (Aron & Aron, 1986) posits that connecting with outgroup individuals (and their groups) enables sharing their perspectives, material/social resources, or identities, reducing bias (Wright et al., 2005). Thus, trust can build a cohesive society across groups through increasing closeness and connection between individuals.

Building Trust

A particularly promising approach to building trust between strangers arises from Interdependence Theory (Kelley & Thibaut, 1978). Historically, social psychologists saw cooperation as the ideal interaction structure for reducing prejudice and intergroup conflict (Allport, 1954; Sherif, 1966). However, according to Interdependence Theory (Kelley &

Thibaut, 1978) as societal trust levels fall, high-stakes interdependence—not just classic cooperation—is needed. Purely cooperative situations are not diagnostic of trust, as actions benefiting a partner also benefit the actor. Mixed-motive (e.g., prisoner’s dilemma) scenarios reveal trust by making partners choose between exploiting (for personal gain) and trusting (at risk of loss) each other. Provided people behave in trustworthy ways, situations involving high-stakes interdependence can advance relationships beyond mere liking to lasting trust and closeness. Once rooted, trust can improve relations even across group boundaries by boosting willingness to rely on outgroup others (Hewstone et al., 2008) and by fostering friendship formation via increased self-disclosure and intimacy (Reis & Shaver, 1988). Focusing on high-stakes interdependence as a mechanism to build trust is a novel approach. Interdependence Theory analyzes the structure of interpersonal situations, claiming trust matters most when individuals must depend upon each other *and* their interests at least partially conflict (Van Lange & Rusbult, 2012). A meta-analysis confirms that trust does predict cooperation more strongly in social dilemmas with larger conflicts of interest (Balliet & Van Lange, 2013).

Using a Social Dilemma to Build Trust

Our research seeks to demonstrate that high-stakes interdependence can build trust and cohesion, even across group boundaries, leading to closer relationships that last. Extending prior findings that *simulating* costly cooperation enabled cross-group trust even after a transgression (Bergsieker, 2012), this work introduces a new dyadic trust-building procedure. Here, we create high-stakes interdependence using the prisoner’s dilemma (PD) game, a well-validated paradigm in behavioural economics and psychology traditionally used to measure—not induce—trust. The PD game puts the interests of two parties in conflict, tempting each to exploit the other rather than cooperate, rendering their responses diagnostic of trust. After an intimate face-to-face

interaction, participants in an iterated PD game almost always cooperate (Bergsieker, 2012) despite sacrificing their own self-interest, allowing trust to build iteratively over time (Axelrod, 1984; Lount et al., 2008). Somewhat counter-intuitively, dilemmas that pit people against each other create a trust-diagnostic situation that lets individuals display and infer trustworthiness more readily than purely cooperative settings without conflicts of interest. Thus, a counterintuitive hypothesis arises:

Placing people—even those from different groups—in structured conflict-of-interest situations enables them to establish enduring trust better than in “safer” cooperative settings.

An iterated PD game can build trust because it elicits more positive interpersonal actions than the people involved initially expect. Three claims underlie this logic:

1. People doubt others' trustworthiness in social dilemmas and avoid potential betrayals.
2. Actual behaviour in high-interdependence situations tends to be trusting due to strong social and moral injunctive norms.
3. More trust is inferred when a partner behaves cooperatively despite a stronger *apparent* temptation to act selfishly arising from the situation.

In social dilemmas, people expect that the majority of strangers will *not* reciprocate their trust (Dunning et al., 2014; Fetchenhauer & Dunning, 2009, 2012). Trust betrayals lead to pain and self-blame, so people prefer risking losses inflicted at random than by a person (Efron & Miller, 2011). However, once immersed in interactions, the majority of people will act as though they trust a stranger, even when they see reciprocity as unlikely (Fetchenhauer & Dunning, 2009, 2012). In mixed-motive situations, people make trusting choices because they feel they “should,” they anticipate anxiety and guilt otherwise, and they feel the choice to trust shows respect for a partner (Dunning et al., 2014). More so than other mixed-motive games (e.g., the ultimatum or

“trust” game; Berg et al., 1995) an iterative PD game can remove much of the “noisiness” of interpersonal interactions: The binary *cooperate/compete* PD game choice reduces uncertainty about how to communicate benevolence to one’s partner and how to detect favourable partner responses in return. For example, after a friendly interaction with a same- or cross-race stranger, trusting choices approached 100% (Bergsieker, 2012).

The juxtaposition of negative expectations about a partner’s likelihood of cooperation and their actual (mostly) cooperative behaviour leads to the inference of trustworthiness. Tendencies to overlook situational, normative influences on behaviour and to draw dispositional inferences (Ross, 1977) lead to inferred trust, especially when the apparent costs of cooperation are high. Given general avoidance of trust-diagnostic situations and cross-group interactions, an outgroup member’s “surprisingly” positive behaviour serves as a powerful learning tool that enables people to view outgroups in a new light.

Overview of Experiments

A series of live-interaction and forecasting experiments (N = 1001) placed same- and cross-race pairs of strangers in situations of either high-stakes interdependence—with motives that partly conflict but still incentivize cooperation—or low-stakes classic cooperation without risk. All studies involved closeness-building tasks with a partner, followed by an iterative PD game. The stakes (high vs. low) of the PD game was our focal manipulation. The high-stakes PD game, referred to as the “risky” condition, included two elements intended to increase psychological (and actual) interdependence between pairs of participants: full information about partners’ choices during the iterative PD game and real monetary payment for one trial (ostensibly selected at random). In the low-stakes “safe” condition both elements were absent: participants did not receive information during the interaction or payment afterwards.

Studies 1-3 contrasted the risky and safe PD games to directly test our hypothesis that inducing interdependence will increase trust in live dyadic interactions between previously unacquainted participants (here, termed “experiencers”). Studies 1-3 also test if the risky (vs. safe) PD game would have a greater effect on increasing trust compared to liking, and that partner cooperation is the mechanism underpinning these effects. Studies 2 and 3 then added a negotiation task to determine whether the hypothesized trust gains generalized to novel interactive contexts.

We also tested the prediction that such trust gains arising from risky interdependence are non-obvious and often underestimated. In Studies 4 and 5, participants (serving as “forecasters”) vividly *imagined* completing the closeness-building tasks and (risky or safe) PD game with a partner before completing many of the same outcomes from the experiencer studies. In these studies, forecasters imagined that their partner was “a typical undergraduate” (Study 4) or a specific person presented via a profile (Study 5).

Studies 1-3 and 5 also varied whether participants were paired with a same- or cross-race (real or imagined) stranger of the same gender, creating a 2 (stakes: “risky” or “safe”) by 2 (dyad race: same- or cross-race) design. Study 4 did not specify partner race, but featured a within-participants extension in which forecasters imagined both risky and safe PD games (counterbalanced), providing an even more conservative test of whether stakes effects are obvious to lay people.

Due to the similarity across the five studies’ methods, manipulation, and measures, we present their method and results in parallel for concision, with differences between studies noted.

Transparency and Openness

All studies were approved by an institutional research ethics board (#18533) and complied with APA ethical standards (e.g., informed consent). This program of research is reported completely, including all experiments conducted, with all measures, manipulations, and exclusions. Secondary measures and analyses are reported in named appendices (each referenced in the main text). Study materials, a de-identified merged dataset for all five studies, and analysis code are publicly available on OSF

https://osf.io/qzrur/?view_only=74c0714a58964f048289083507bb58ba).

CHAPTER 2: METHOD

All participants either completed a series of tasks with another student as “experiencers” in Studies 1-3 or imagined doing so as “forecasters” in Studies 4 and 5 (see study features in Table 1). For clarity, we use the term “experiencers” or “forecasters” (which were not shown to participants) when describing procedures unique to one role and “participants” for procedures shared across both roles.

Table 1: Study features

Study	Role	Design	Partner race	Location	Negotiation	PD matrix
1	Experiencer	Between	White/East Asian	In-lab	—	7/3 or 9/1
2	Experiencer	Between	White/East Asian	In-lab	Yes	7/3
3	Experiencer	Between	White/East Asian	In-lab	Yes	7/3
4	Forecaster	Within	Unspecified “typical undergraduate”	In-lab	—	7/3 or 9/1
5	Forecaster	Between	White/East Asian participant profile	Online	—	7/3 or 9/1

Note. In Study 5, participant race was manipulated using descriptive profiles. Matrix refers to the possible PD game cash payout, presented as the ratio of mutual cooperation/mutual defection (7/3 or 9/1).

Participants and Recruitment

We recruited undergraduate students at a large Canadian public university into a study on “interpersonal processes” (Studies 1-4) or “predicting interpersonal processes” (Study 5).

Eligibility

Based on a pilot study testing our PD game paradigm among different populations (Appendix A), the largest reported trust gap in this population emerged between White and East Asian students. Thus, for Studies 1-5, we recruited only individuals who self-identified during pre-screening as White or East Asian to participate for partial course credit or payment (plus a bonus of up to \$10 in the risky condition). Because social value orientation (SVO; e.g., pro-social, individualist, or competitor) has been shown to influence cooperative behaviour (Smeesters et al., 2003), we recruited only individuals whose SVO could be reliably classified

(e.g., approximately 81% of the departmental participant pool for Study 1) based on standard scoring procedures (Van Lange et al., 1997).

Demographics

After exclusions (detailed in Table 2), our final sample consisted of 1001 participants: 578 experiencers and 423 forecasters; 334 men and 667 women; 424 East Asian and 580 White participants; with a median age of 19 (for demographics by study, see Table 3). Nonparametric tests revealed no differences in gender, race, or socioeconomic status between experiencers and forecasters (χ^2 s < 4.93, $ps > .295$). We randomly assigned 500 participants to the safe condition and 501 participants to the risky condition.

Table 2: Exclusions by Study

	Study 1	Study 2	Study 3	Study 4	Study 5	Total
All participants	208	194	282	118	397	1177
Exclusions						
Suspicion	10	11	24	2	21	68
Technical issues	8	8	1	4	1	22
Failed instructions	4	2	0	0	8	14
Misperceived race	4	1	1	0	27	33
Prior acquaintance	2	0	2	0	0	4
Experimenter error	0	6	0	3	0	9
Ineligible/incomplete	0	11	1	0	26	16
Outlier	0	1	9	0	0	10
Total exclusions	28	40	38	9	61	176
Final N	180	154	244	109	314	1001
Risky PD game	51%	56%	47%	47%	50%	50%
Safe PD game	49%	44%	53%	53%	50%	50%
Viewed standard 7/3 matrix	58%	100%	100%	43%	60%	74%

Note. Ineligible participants either completed a related study in the past or were not White or East Asian.

Table 3: Participant Demographic Information by Study

Demographics	Study 1	Study 2	Study 3	Study 4	Study 5	Total
Median age	19	19	19	20	19	19
Participant race						
White	62%	58%	55%	54%	59%	58%
East Asian	38%	42%	45%	46%	41%	42%
Gender						
Men	23%	42%	34%	19%	39%	67%
Women	77%	58%	66%	81%	61%	33%
Socioeconomic class						
Working/lower-middle	15%	17%	10%	13%	19%	15%
Middle	61%	47%	52%	54%	51%	53%
Upper-middle/upper	24%	36%	37%	33%	31%	32%
Social value orientation						
Pro-social	65%	69%	64%	64%	57%	63%
Individualist	27%	22%	32%	31%	31%	29%
Competitor	8%	9%	4%	5%	12%	8%

Note. Two participants in Study 3 were missing social value orientation. In each study 5-14% of participants did not report socioeconomic status.

Procedure

Study sections were ordered as follows: Getting acquainted, trust-diagnostic task, negotiation (Studies 2 & 3), end-of-study questionnaires. Each section outlines materials provided to participants, activity descriptions, and relevant measures. For more detailed descriptions, see the *Materials* section.

Research Timeline and Setting

We ran Studies 1 and 4 concurrently, and Studies 1-4 were conducted in lab, whereas Study 5 was online via Qualtrics. In-lab participants were seated in separate rooms when completing surveys via Qualtrics or the PD game on networked computers using Z-tree software for experiments (Fischbacher, 2007).¹

¹ The first 12 dyads (14% of Study 1) and all pilot study participants used a paper-based version of the PD game that involved sliding cards with their choices under the door of their individual rooms, which the experimenter then passed to the other partner. Game format (paper-based vs. online) did not moderate any reported results.

Stopping Rules and Statistical Power. Sample sizes were based on the maximum number of participants we could recruit during specific terms (with at least 100 participants per study). No data analysis was conducted until data collection for each study had ended. Post-hoc sensitivity analysis (using G*Power, Faul et al., 2007) with 80% power ($\alpha = .05$) based on our sample size of 1001 yielded a minimum detectable effect size for mean differences of $d = 0.18$.

Experimenter Session Log. A White or East Asian female experimenter ran each session and recorded the date and time; condition assignment; participants' first names, apparent gender and race, and punctuality; experiencers' choices in the PD game (needed for payment); technical issues, suspicions expressed during debriefing; anomalies; and pay for experiencers.

Getting Acquainted

We included a getting-acquainted phase to foster a positive interpersonal connection conducive to subsequent cooperation, even enabling recovery to maximal rates of cooperation following defection in iterative cross-race PD games (Bergsieker, 2012). Although the goals of this phase were social, we described activities as “tasks” based on past experimental (Babbitt & Sommers, 2011) and meta-analytic (Toosi et al., 2012) evidence that framing cross-race interactions as structured tasks rather than free-form social interactions reduces cognitive depletion and performance gaps between cross-race and same-race dyads.

Introductions. Participants in Studies 1-4 came to the lab for a study that involved “joint activities, discussion, and decision making” with another student (whoever had signed up for the same session; if no partner was available for a scheduled session, participants were assigned to Study 4 in the role of forecasters). In Studies 1-3, the experimenter first confirmed that pairs of experiencers were strangers, then seated them facing each other at a table in the interaction room. In Studies 4 and 5, forecasters participated individually and were instructed to “vividly imagine”

completing the following tasks with “a typical University of Waterloo undergraduate student of your gender” (in Study 4) or with an ostensible partner whose profile they viewed (in Study 5). In Study 5, forecasters viewed the gender-matched profile of a White or East Asian partner (thus varying dyad race, as in Studies 1-3), after first answering demographic questions to build their own personal profile (to bolster the believability of the “partner” profile).

Closeness-Inducing Tasks. Experiencers then completed (and forecasters imagined completing) three tasks designed to increase interpersonal closeness: identifying similarities, split-face drawing, and discussion of “Fast Friends” prompts with partners. Next, participants completed an instructional manipulation check and questionnaires about their task experiences.

Trust-Diagnostic Phase

Before participants learned any details of the upcoming “joint decision-making task” (the PD game), they reported their initial (pre-game) feelings about their relationship with their partner (including trust, liking, and closeness).

All participants were randomly assigned to either a “risky” (high-stakes) or “safe” (low-stakes) version of a PD game. Experiencers were always assigned to the same condition as their partner. Written instructions in each study varied the stakes via task-based payment and information about partner choices (to reinforce the actual stakes, the experimenter told experiencers in the risky condition that they could earn a bonus between \$0 and \$10 in a joint decision-making task). Participants had to correctly answer questions about the pay-offs prior to making predictions or decisions in the game (repeated failure informed participant exclusions).

Next, participants completed (or imagined completing) a risky or safe 15-round iterative PD game with their partner. After the PD game, participants individually completed questionnaires about their task experiences and post-game interpersonal impressions.

Negotiation Task (Studies 2 & 3)

Experiencers then returned to the interaction room for a 10-minute negotiation task, followed by questions about their negotiation experience, and their partner.

Additional Tasks

Experiencers completed additional measures related to their social network in all studies, as well as a “current issues task” in Studies 1 and 2 (also completed by forecasters) or a “decision prediction task” in Study 3 (see Appendix B). We included these measures to explore whether completing the risky PD game with a cross- rather than same-race partner might lead to more cross-race (vs. same-race) social network integration, willingness to discuss race-related (vs. race-neutral) issues, and expected cooperation from future cross-race (vs. same-race) partners in the PD game. However, no reliable condition effects emerged on these variables, so they are not discussed further.

End-of-Study Measures

Next, participants completed individual difference measures of general trustfulness and, in Studies 2, 3, and 5b, social desirability (Fischer & Fick, 1993). Participants answered questions about manipulation checks, their demographic background, and their engagement level. Participants were then individually thanked, debriefed, and compensated.

Materials

Except where specific studies are noted parenthetically, materials and measures were included in all studies (for scale items and reliabilities, see Appendix C).

Partner Profiles (Study 5)

To more closely mirror the nature and content of the dyadic lab tasks completed with a stranger, forecasters in Study 5 received a partner profile based on a specific prior participant’s

responses. In addition to increased realism, this approach reduces the chances of participants envisioning a close friend when asked to imagine a “typical undergraduate” (as in Study 4).

The first wave of data collection (Study 5a) used gender-specific standardized profiles based on characteristics often shared by experiencers during the similarities task (e.g., Psychology major, past travel to California, enjoys working out). To convey apparent race, profiles included a first name that was stereotypically White (Emily, Ethan) or East Asian (Hao, Yan) and respective birth country of Canada or China. To boost believability, participants first saw their own profile (based on answers to demographic and background questions), with the option to revise it for potential use in future studies. Next, in a “yoked” design, the second wave of data collection (Study 5b) used 40 of these profiles (provided by consenting Study 5a participants), to further increase realism and comparability to the experiencer version of the study. In Study 5b, all participants viewed a gender-matched partner profile of a White or East Asian previous participant.²

Getting Acquainted: Tasks and Measures

Closeness-Inducing Tasks. Participants completed (or imagined completing) three tasks with their partner. In the “Venn Diagram” task, participants had 3-5 minutes to identify and list their similarities in the (large) overlapping portion of a Venn diagram. Next, in the “face-drawing” task, participants had 3 minutes to draw half of their own face and half of their partner’s face on a paper folded in half, to visually reflect merging of self and other (Aron et al., unpublished manuscript). In the third task, participants had 5 minutes to discuss four questions (of their choosing) from a list of eight self-disclosure questions (e.g., “What do you value most in a friendship?”) adapted from established “Fast Friends” procedures for generating closeness

² Study 5b stated “racial/ethnic group” explicitly, rather than relying on first name and birth country to imply race.

between strangers (Aron et al., 1997). We used page timers to ensure that forecasters had sufficient time to vividly imagine each task. After the discussion task, participants rated how interesting, difficult, and enjoyable each of the three tasks was (or would be), from 1 (*Not at all*) to 7 (*Extremely*).

Relational Measures: Trust, Liking, Closeness (Studies 1-3). Experiencers rated their perceptions of trust, liking, and closeness toward their partner using continuous sliders from *Not at all* to *Extremely* in 100 increments. (The slider scale did not display numerical labels and is scored from 1 to 7 to aid comparisons across measures.) For concision, we created trust, liking and closeness composites: Trust (two items) combined reliability and trust items, liking (two items) combined enjoyment and liking, and closeness combined feeling close to partner and knowing them well. Testing the individual trust, liking, and closeness items yields identical conclusions. Secondary slider measures of similarity and friendship interest were also included.

To assess self-other overlap (a specific form of closeness), we used the Inclusion of Other in Self scale (IOS; Aron et al., 1992) that instructed participants to “select the pair of circles below that you feel best represents how close you are to your partner.” Experiencers chose between seven pairs of circles, ranging from completely non-overlapping to almost completely overlapping. Experiencers completed the slider and IOS measures twice: shortly before and again after the PD game.

Trust-Diagnostic Phase: Tasks and Measures

Instructional Manipulation Check. As our central manipulation of PD game stakes relied on written instructions, participants were first given an instructional manipulation check (IMC), a validated tool for prompting inattentive participants to follow instructions, increasing data reliability (Oppenheimer et al., 2009). Participants read a paragraph with embedded

instructions followed by a rating scale and a textbox. To pass, participants had to leave the rating scale blank and type “I have read the instructions” in the textbox. Repeated IMC failure informed our decisions to exclude participants from analysis.

PD Game. Participants completed (or imagined completing) 15 trials of the PD game, choosing between either cooperation (“X”) or defection (“Y”) each round. The combined choices for each round determined earnings for both parties. After viewing a PD payoff matrix, participants verified their comprehension by indicating their expected earnings in each of the four possible scenarios (both cooperate, partner defect, self defect, both defect). To encourage participants to begin with cooperative choices, we showed participants a “Tutorial on Cooperation” (adapted from Murnighan, 1991, pp. 13–27) on the benefits of cooperation across repeated trials in mixed-motive tasks (see also Bottom et al., 2002; Lount et al., 2008).

PD Game Matrix. We presented participants with a square grid (or “matrix”) of payoffs based on recommendations from research formally quantifying the extent to which individuals’ motives converge or diverge within the PD game (Kelley & Thibaut, 1978). Setting equal values for *temptation* (i.e., payoff for sole defection minus mutual cooperation), *risk* (i.e., payoff for joint defection minus sole cooperation), and *gain* (i.e., payoff for mutual cooperation minus mutual defection) in a 1:1:1 ratio (index of cooperation $k = .33$; Rapoport, 1967) considered optimal (Kelley & Thibaut, 1978) and normative in PD game research (e.g., typical k range = .33-.50 in studies reviewed by Balliet & Van Lange, 2013). For a \$0-\$10 range, these payoffs correspond to \$6.67 each for mutual cooperation, \$3.33 each for mutual defection, or \$10 for the defector and \$0 for the cooperator if choices differed. We rounded these amounts to the nearest dollar to increase perceptual fluency and payment ease, so our participants could receive \$7, \$3, \$10 or \$0 (temptation-risk-reward ratio = 1:1:1.33; $k = .40$). Setting equal values for *temptation*

(i.e., payoff for sole defection minus mutual cooperation), *risk* (i.e., payoff for joint defection minus sole cooperation), and *gain* (i.e., payoff for mutual cooperation minus mutual defection) in a 1:1:1 ratio (index of cooperation $k = .33$; Rapoport, 1967) is considered optimal (Kelley & Thibaut, 1978) and normative in PD game research (e.g., typical k range = .33-.50 in studies reviewed by Balliet & Van Lange, 2013). For a \$0-\$10 range, these payoffs correspond to \$6.67 each for mutual cooperation, \$3.33 each for mutual defection, or \$10 for the defector and \$0 for the cooperator if choices differed. We rounded these amounts to the nearest dollar to increase perceptual fluency and payment ease, so our participants could receive \$7, \$3, \$10 or \$0 (temptation-risk-reward ratio = 1:1:1.33; $k = .40$).

The first experimenter study (and both forecaster studies) varied the PD game matrix to explore whether \$9, \$1, \$10, and \$0 payoffs (i.e., the “9/1 matrix”) that reduced *temptation* to only \$1 (\$10 - \$9), would attenuate the hypothesized trust-inducing effects of the PD game with the optimal index of cooperation (the previously described “7/3 matrix”). Because payoff matrices did not significantly influence focal outcomes, the later experimenter studies used the 7/3 matrix exclusively and we collapsed across matrix types in all analyses (for matrix-related analyses, see Appendix D).

Stakes Manipulation. In the risky condition, participants were told that they would each (a) receive the actual dollar amount they earned in a randomly selected trial (in fact, all task-based payments were based on the 14th trial), and (b) see their partner’s choices and earnings. After each trial (except the last), these participants saw an outcome screen with each partner’s choice (“X” or “Y”) and earnings. Right before the 15th trial, participants were informed that the final trial would be secret and their partner would not know their choice. Conversely, in the safe condition, participants were told that all earnings were hypothetical: They would not receive

actual money for the task or learn their partner's choices after each trial. These participants saw no outcome screens. As forecasters instead imagined completing the PD game, rather than completing it, they could not earn additional remuneration in the risky condition.

PD Game Predictions. Participants made a series of pre-game predictions about cooperation (described as “choosing X”) or defection (“choosing Y”) in the PD game, and their confidence in each prediction, rated from 1 (*not at all*) to 5 (*completely*).³ Specifically, participants were asked to predict first-round cooperation (vs. defection) by their partner, a “typical undergraduate” (experiencers only), and themselves (forecasters only); final-round cooperation by their partner and themselves (forecasters only); and whether they would respond to a partner's defection with defection (forecasters only). For brevity and ease of interpretation, these predictions and their respective confidence scores were combined into “confidence of cooperation” composites, where if participants predicted defection they were scored as a 0, and otherwise their confidence score in their cooperation prediction was used. All participants predicted the likelihood of their partner never defecting from 0% to 100%.

PD Game Cooperation. For each of the 15 rounds, experiencers could cooperate or defect. We scored cooperations as a 1 and defections as a 0, then computed the average rate of cooperation across 15 rounds. We also created a binary “full cooperation” variable to identify participants who cooperated in all 15 rounds (scored 1) from those who did not (scored 0). Forecasters predicted how frequently they or their partner would defect across all 15 rounds, with seven binned response options (*never, 1-3 times, 4-6 times, half the time, 9-11 times, 12-14 times, always*). To facilitate direct comparisons to experiencers' actual cooperation, we

³ Studies 1, 4 and 5a used a scale from 0-100% for confidence in predictions, however we switched to a 1-5 to avoid interpretational difficulties for Studies 2, 3 and 5b. All analyses rescaled all predictions to 1-5 to combine across studies.

converted these estimates to the same numeric scale for average rate of cooperation using the midpoint of each response bin (e.g., defecting “1-3 times” was coded as cooperating in 13 out of 15 rounds) and also created a binary variable encoding whether forecasters predicted full cooperation (i.e., their partner “never” choosing Y; scored 1) or not (scored 0).

PD Game Reactions. Immediately after completing (or imagining) the PD game, experiencers in Studies 2 and 3 indicated their affective reactions: feeling *grateful*, *resentful*, *surprised*, or *relieved*, each rated from 1 (*Not at all*) to 7 (*Very much*). Based on reliability analyses, we combined gratitude and relief into a composite. In all studies, experiencers then reported (and forecasters predicted) the extent to which they cared about their partner’s choices, were tempted to defect (choose “Y”) at least once, and were worried their partner might defect at least once, again rated from 1 (*Not at all*) to 7 (*Very much*).

Post-Game Trust. Trust is central to this line of research, so we measured it in several ways: the continuous sliders previously described (covarying for pre-game trust using this more gradated measure), face-valid single-item measures, perceptions of trust impact and change, and a detailed scale designed to assess interpersonal trust, cultural trust, and liking after a dyadic interaction (Bergsieker, 2012). However, this final 28-item scale was too long for use in Study 4 (where forecasters considered four versions of the PD game in succession) and yielded results largely redundant with those in the main text, so its results are reported separately (see Appendix E).

Single-Item Measures. Soon after the PD game, participants responded to single-item measures of the extent to which they trusted, liked, and could rely on their partner, rated from 1 (*Not at all*) to 7 (*Very much*). Perceived impact measures assessed participants’ belief that the PD game influenced their trust in and closeness to their partner.

Perceived Trust and Relationship Change. Participants rated the extent to which the PD game changed the quality of the relationship with their interaction partner using a 7-point semantic differential scale (with the midpoint described as meaning “no change” or “neither”). Items included: *Less trusting – More trusting*, *Stronger – Weaker* (reverse scored), *Less close – More close*, *Better – Worse* (reverse scored), and *Less pleasant – More pleasant*.

Negotiation Task (Studies 2 & 3)

Back in the interaction room, experiencers engaged in a brief negotiation task with integrative potential: A modified version of the *Kukui Nuts* simulation (Kopelman & Berkel, 2012).⁴ Experiencers were instructed to negotiate for 10 minutes as representatives from two companies (after taking 5 minutes to read individual role instructions). In Study 3, the negotiation was video recorded (and participants’ behaviour was coded, however few effects emerged; for details see Appendix F).

Post-Negotiation Questionnaires. Experiencers rated the extent to which they were satisfied with communication during the negotiation, their own negotiation outcomes, their partners’ negotiation outcomes, and their relationship with their partner after the negotiation, from 1 (*not at all*) to 7 (*perfectly/a great deal*). Then, experiencers rated the perceived change in their relationship caused by the negotiation using the same response scale as *Perceived Change* section earlier. After that, experiencers answered questions about trust, engagement, information sharing, understanding, satisfaction, and feeling manipulated during the negotiation, on a scale from 1 (*strongly disagree*) to 7 (*strongly agree*). These items were used to create a six two-item composite scores (averaging across items): information sharing, satisfaction with negotiation

⁴ In the source task, some prior participants found an integrative solution for sharing all nuts (based on each party’s needs). Among our participants, who had less time to negotiate, this non-zero-sum solution was too rare to analyze. More details on the negotiation task are found in Appendix B.

outcomes, partner treatment during negotiation, comfort during negotiation, mutual understanding, and partner warmth and competence, and one three-item composite score for post-negotiation trust.

Individual Difference and Demographic Measures

Participants completed a 5-item general trustfulness measure (Yamagishi, 1986) on a scale from 1 (*Strongly Disagree*) to 7 (*Strongly Agree*). Experiencers also completed the 6-item short form (see Fischer & Fick, 1993) of a social desirability measure (Crowne & Marlowe, 1960) with *True* and *False* as response options either in the department-wide mass testing survey or within the study (in Studies 2 and 3). Participants reported their age, gender⁵, year in school, primary and perceived racial group (i.e., “How do you think most people would describe your primary racial/ethnic group?”), and their socioeconomic status.

Manipulation Checks, Engagement Checks, and Suspicion Probes

Manipulation checks asked participants to indicate their partner’s gender, race, current year in school, and how often they and their partner defected (chose “Y”) during the PD game. Participants reported their level of engagement (i.e., distraction, seriousness) and whether they were suspicious about the study, then a verbal funnel debriefing also probed for suspicion.

Analytic Approach

Given the dyadic design of Studies 1-3, all analyses of experiencer and aggregate data used multi-level models with participants nested within dyads to account for nonindependence (Kenny et al., 2006). In mega-analyses, forecasters were considered nested within dyads, however as forecasters participated alone their ‘dyads’ only contained themselves. We used the MIXED procedure in SPSS to estimate these models with compound symmetry for

⁵ This measure conflated gender and sex: Participants could select only “Male” or “Female.”

indistinguishable dyads (although stakes varied only between dyads, race could vary between or within dyads), as well as the lme4 and mediate packages in R for multi-level mediation analyses.

To enable more accurate estimation and concise reporting of results, we used integrative data analysis (Curran & Hussong, 2009), also known as “mega-analysis” (Costafreda, 2009). Mega-analysis of participant-level data pooled across experiments is the gold standard method for integrating study-level statistics (Cooper & Patall, 2009; Costafreda, 2009; Curran & Hussong, 2009; Sung et al., 2014) and especially suitable when experimental designs are highly similar across studies. This mega-analytic approach also enables higher-powered tests of moderation (e.g., by dyad race or social value orientation). Thus, to minimize repetition and highlight importance and consistency of effects, we mega-analyzed our data whenever possible.

Core Model

Our core model included the following effects-coded predictors: PD game stakes (safe = -0.5, risky = 0.5), dyad race (cross-race = -0.5, same-race = 0.5), and their interaction.⁶ Using these unweighted fractional effects codes aids interpretation by making the unstandardized *bs* equivalent to the mean difference between the two conditions. When comparing two conditions, we also report Cohen’s *ds*, with each operative effect size (see Judd et al., 2017) calculated based on Rosenthal and Rubin (2003; Equation 3).

Model Modifications and Extensions

When directly testing whether the stakes effects varied between experiencers and forecasters, an additional model added role (experiencer = 0.5, forecaster = -0.5) and its interaction with stakes. To test whether stakes effects varied based on SVO, gender, general trustfulness, or social desirability, supplemental analyses extended the core model to include

⁶ We retained dyad race in the core model (despite yielding few effects) due to its a priori theoretical importance, and because its lack of interaction with stakes addresses intervention efficacy in both same- and cross-race contexts.

each potential moderator in turn (see Appendix G). In general, we probed any interactions by appropriately recoding categorical predictors or rescaling continuous predictors within the full sample (Aiken & West, 1991), but because data from forecasters and experiencers came from separate studies, follow-up analyses probing interactions between stakes and role involved estimating simple effects of stakes for forecasters and experiencers separately.

For the post-game slider variables, we covaried for experiencers' pre-game ratings (made on identical measures) to more precisely isolate effects of the PD game from pre-existing variability in participants' trust, liking, and closeness (omitting these covariates yielded identical conclusions). For Study 4 forecasters, primary analyses focus on the first scenario viewed (within-participant analyses using repeated-measures ANOVA revealed similar effects, see Appendix H).

CHAPTER 3: RESULTS

The main text presents mega-analytic results, combining all five studies when applicable and probing role-specific effects of the stakes manipulation within experiencers (Studies 1-3) and forecasters (Studies 4 & 5), separately. Although some dependent variables (e.g., those related to the negotiation task) were not collected in every study, the mega-analysis includes all available data across studies (for study-by-study results, where available, see the Appendix I). Results are organized into five sections respectively testing whether effects of PD game stakes (1) emerged on key outcomes for experiencers, (2) emerged on these outcomes for forecasters or varied by role (experiencer vs. forecaster), (3) affected trust- more than liking-related outcomes, (4) extended to a novel context negotiation task, and (5) were mediated by cooperation rates.

Minimal study-level clustering emerged: Intraclass correlations (ICCs) testing clustering at the study level for each reported measure were so small that many models did not converge (in models that did converge, ICCs ranged from .001 to .07), so study-level variability was not analyzed further. As expected, responses of individuals within dyads showed substantial clustering (for dyad ICCs of key variables, see Appendix J).

Stakes Effects on Experiencers

First, we report the effects of stakes on key outcomes for experiencers: cooperation, trust, and liking and other relationship indicators. We predicted that experiencers in the risky (vs. safe) condition would cooperate more frequently, report higher trust, and higher liking and other relationship indicators.

Consistent with random assignment to PD game condition, no significant condition differences emerged in experiencers' pre-game trust, liking, or closeness (all t s < 1.63, all p s > .013), measured immediately *prior* to the PD game instructions about its safe or risky stakes.

Did the Game Stakes Affect Predicted and Actual Cooperation for Experiencers?

There was no significant difference between risky (97%) and safe (97%) for experiencer predictions of their partner cooperating in the first round ($\chi^2 = 0.07, p = .798$; Table 4). More experiencers in the risky (82%) condition than the safe (73%) condition expected a typical undergraduate to cooperate in the first round ($\chi^2 = 6.41, p = .011$). More experiencers in the risky (99%) condition than the safe (96%) condition cooperated in the first round ($\chi^2 = 6.57, p = .010$). Finally, more experiencers in the risky (94%) condition than the safe (71%) condition cooperated for all 15 rounds ($\chi^2 = 53.54, p < .001$).

Table 4: Binary Cooperation Measures by Stakes and Role

Outcome	N	Experiencers			Forecasters			Role		
		Overall	Safe	Risky	χ^2	Overall	Safe	Risky	χ^2	χ^2
Predicted partner coop R1	995	97%	97%	97%	0.07	89%	86%	92%	3.26 [†]	23.25^{***}
Predicted undergrad coop R1	681	77%	73%	82%	6.41[*]	87%	81%	94%	4.15[*]	5.36[*]
Round 1 cooperation	999	98%	96%	99%	6.57^{**}	90%	89%	91%	0.86	27.96^{***}
Rate of full cooperation	1001	82%	71%	94%	53.54^{***}	35%	40%	30%	4.67[*]	231.31^{***}

Note. R1 = Round One. Coop = cooperation. Boldface indicates significant effects. Cooperated Round 1 and full cooperation rate are actual behaviours for experiencers, but predictions for forecasters. *** $p < .001$. ** $p < .01$. * $p < .05$. † $p < .1$.

Stakes impact experiencers’ predictions and inferences about cooperation in the PD game, as well as their own actual behaviour in the PD game (see Table 5, left-most columns). Experiencers in the risky (vs. safe) condition crucially did *not* have greater confidence that their own partner would choose to cooperate ($t < 1$), however they were more confident that a typical undergraduate would cooperate ($b = 0.29, d = 0.25, p = .032$).

During the actual PD game, experiencers in the risky condition cooperated more overall ($b = 0.09, d = 0.89$), cared more about their partner’s choices in the PD game ($b = 1.24, d = 1.01$), felt less tempted to defect ($b = -0.85, d = -0.54$), and reported greater gratitude and relief ($b = 1.76, d = 1.70$) than those in the safe condition (all $ps < .001$).

Table 5: Continuous Cooperation Measures by Stakes and Role

Outcome	N	Experiencers				Forecasters				Stakes × Role t
		Safe Mean (SD)	Risky Mean (SD)	Stakes t	d	Safe Mean (SD)	Risky Mean (SD)	Stakes t	d	
Cooperation										
Confidence in partner Round One cooperation	979	3.37 (1.10)	3.40 (1.05)	0.38	0.04	2.93 (1.49)	3.11 (1.29)	1.19	0.12	0.90
Confidence in undergrad Round One cooperation ^a	664	2.24 (1.63)	2.52 (1.44)	2.15*	0.25	2.95 (1.60)	3.29 (1.16)	1.28	0.25	0.19
Probability partner cooperates 100%	998	61% (25%)	61% (26%)	0.27	-0.03	50% (32%)	49% (32%)	0.04	0.00	0.11
Own cooperation rate	997	90% (20%)	99% (3%)	7.54***	0.89	79% (0.28)	76% (27%)	1.10	-0.11	4.09***
Partner cooperation rate	997	90% (20%)	99% (3%)	7.43***	0.90	72% (24%)	69% (25%)	1.40	-0.14	5.11***
PD game reactions										
Cared about partner's choices	1001	3.65 (1.67)	4.89 (1.66)	8.68***	1.01	4.11 (1.93)	4.64 (1.65)	2.84**	0.28	3.20**
Tempted to defect	1001	3.30 (2.35)	2.45 (1.86)	4.53***	-0.54	3.92 (2.07)	4.10 (2.18)	0.77	0.08	3.64***
Worried partner would defect	1001	3.14 (1.79)	2.97 (1.60)	1.17	-0.14	4.22 (1.87)	4.29 (1.77)	0.23	0.02	0.97
Gratitude and relief ^b	398	3.45 (1.52)	5.22 (1.23)	12.00***	1.70	—	—	—	—	—

Note. Boldface indicates significant effects. Confidence in cooperation ranged from 0 (predicted defection) to 5 (predicted cooperation and is completely confident in that prediction). Reactions to the PD game were measured on a seven-point scale. ^a Confidence in typical undergraduate Round One cooperation was not measured in Study 5. ^b Gratitude and relief were only measured in Studies 2 and 3.

*** $p < .001$. ** $p < .01$. * $p < .05$. † $p < .1$.

Did the Game Stakes Impact Trust and Relationship Outcomes Among Experiencers?

Stakes impacted experiencers' trust on every operationalization included (see Table 6).

Experiencers who completed the risky (vs. safe) PD game reported higher ratings on all four focal trust measures (all $ps < .001$), single-item trust ($b = 1.21$, $d = 1.46$), willingness to rely on their partner ($b = 1.06$, $d = 1.17$), PD game impact on trust ($b = 1.97$, $d = 1.64$), and perceived increase in trust ($b = 1.39$, $d = 1.96$).

Turning from trust to other relationship dimensions, experiencers in the risky (vs. safe) condition showed smaller (but still substantial) benefits (all $ps < .001$). Completing the risky (vs. safe) PD game led to greater liking ($b = 0.61, d = 0.79$), and greater perceived increase in pleasantness ($b = 1.40, d = 2.10$). Similarly, participants in the risky (vs. safe) PD game perceived increase of ($b = 1.06, d = 1.49$) and reported a greater impact on ($b = 1.51, d = 1.26$) closeness in the relationship. Completing the risky (vs. safe) PD game additionally led participants to perceive more strength in the relationship ($b = 1.01, d = 1.39$), and that the relationship was better ($b = 1.17, d = 1.62$), afterwards. Bivariate correlations emerged between focal variables and secondary variables within each stakes condition for experiencers and forecasters (see Appendix K).

Table 6: Stakes Effects on Trust and Relationship Outcomes

Outcome	<i>N</i>	Experiencers				Forecasters				Role × Stakes <i>t</i>
		Safe Mean (<i>SD</i>)	Risky Mean (<i>SD</i>)	Stakes <i>t</i>	<i>d</i>	Safe Mean (<i>SD</i>)	Risky Mean (<i>SD</i>)	Stakes <i>t</i>	<i>d</i>	
Trust										
Trust (single-item)	970 ^a	4.49 (1.19)	5.70 (1.05)	12.30 ^{***}	1.46	4.18 (1.39)	4.29 (1.43)	0.74	0.07	6.54 ^{***}
Reliance (single-item)	970 ^a	4.31 (1.30)	5.37 (1.22)	9.88 ^{***}	1.17	3.93 (1.38)	3.91 (1.55)	0.08	-0.01	6.00 ^{***}
Trust (PD game impact)	970 ^a	3.37 (1.79)	5.34 (1.49)	14.09 ^{***}	1.64	4.40 (1.62)	4.50 (1.59)	0.47	0.05	8.69 ^{***}
More trusting (perceived change)	1001	4.28 (1.04)	5.66 (0.96)	16.30 ^{***}	1.96	4.02 (1.28)	4.15 (1.26)	1.18	0.11	8.53 ^{***}
Liking										
Liking (single-item)	970 ^a	5.16 (1.06)	5.77 (1.00)	6.60 ^{***}	0.79	4.36 (1.29)	4.43 (1.39)	0.67	0.07	3.25 ^{**}
More pleasant (perceived change)	1001	4.25 (0.92)	5.65 (0.96)	17.32 ^{***}	2.10	4.06 (1.22)	4.21 (1.23)	1.39	0.14	8.87 ^{***}
Relationship										
Closeness (PD game impact)	970 ^a	3.08 (1.63)	4.59 (1.56)	10.92 ^{***}	1.26	4.06 (1.63)	4.08 (1.54)	0.16	0.02	6.98 ^{***}
Closer (perceived change)	1001	4.17 (0.94)	5.23 (1.01)	12.55 ^{***}	1.49	4.07 (1.17)	4.22 (1.07)	1.48	0.14	6.59 ^{***}
Stronger (perceived change)	1001	4.27 (0.93)	5.28 (1.04)	11.67 ^{***}	1.39	4.17 (1.15)	4.19 (1.18)	0.33	0.03	6.96 ^{***}
Better (perceived change)	1001	4.37 (0.93)	5.54 (1.07)	13.46 ^{***}	1.62	4.27 (1.16)	4.27 (1.14)	0.19	0.02	8.29 ^{***}

Note. Boldface indicates significant effects. All measures were on a seven-point scale. ^a 31 participants in study 4 had their single-item scale points scrambled, and are excluded from these analyses.

*** $p < .001$. ** $p < .01$. * $p < .05$. † $p < .1$.

How Did Forecasters' Predictions Compare to Experiencers' Reports?

Next, we investigated differences between forecasters and experiencers, reporting role main effects and the interaction between stakes and role on cooperation rates and key outcomes. We predicted that the unique benefits of inducing interdependence in the risky condition would be underappreciated in forecasters, and that participants would underestimate the prevalence of cooperation overall, and the impact of stakes on cooperative behaviour.

Including participants' role as an experiencer or a forecaster in our mega-analytic model allowed us to formally test the differences of stakes effects on key outcomes for experiencers and forecasters. Significant role main effects emerged such that relative to forecasters, experiencers were more confident that their partner would cooperate in the first round ($b = 6.07, p = .001, d = 0.26$), but less confident that a typical undergraduate ($b = 20.09, p < .001, d = 0.51$) would cooperate in the first round, and reported less temptation to defect or concern about their partner defecting ($bs > 1.13, ps < .001, ds > 0.61$). Forecasters also predicted less cooperation from themselves ($b = 0.17, p < .001, d = 1.00$) and their partners ($b = 0.24, p < .001, d = 1.60$) than experiencers and their partners actually cooperated. Except for perceived trust impact, experiencers (vs. forecasters) reported higher trust on all other focal trust measures ($bs > 6.40, ps < .001, ds > 0.79$). Experiencers (vs. forecasters) also report higher single-item liking ($b = 1.07, p < .001, d = 1.10$), and more positive relationship change ($b = 0.67, p < .001, d = 0.88$), however also lower perceived closeness impact ($b = 0.23, p = .029, d = 0.16$).

We also compared experiencers' and forecasters' binary predictions about cooperation to actual cooperation behaviour (Table 4). Overall, more experiencers (97%) than forecasters (89%) expected their partner to cooperate in the first round ($\chi^2 = 23.25, p < .001$), but fewer experiencers (77%) than forecasters (87%) expected a typical undergraduate to do so ($\chi^2 = 4.15, p = .021$). Experiencers (98%) not only cooperated more in the first round than forecasters (90%) predicted ($\chi^2 = 27.96, p < .001$), experiencers (82%) also cooperated fully in all rounds much more frequently than forecasters (35%) predicted ($\chi^2 = 231.31, p < .001$).

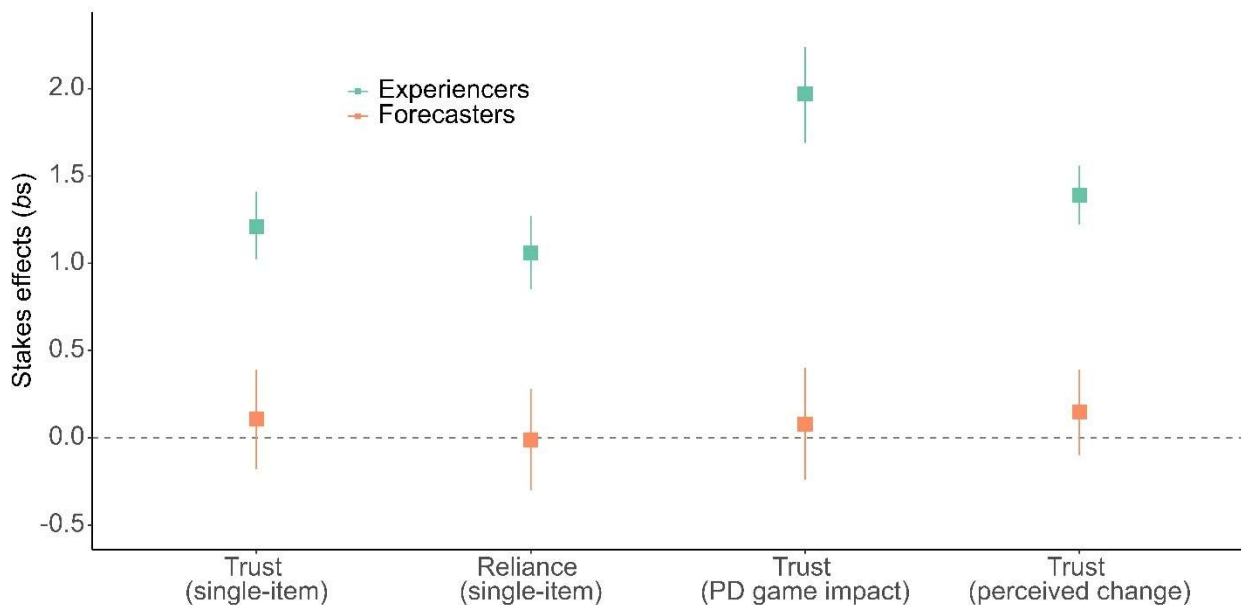
Stakes by Role Moderation: Did Forecasters Foresee these Stakes Effects?

Tests of moderation by role for cooperation variables (see right-most column of Table 5) revealed larger stakes effects for experiencers on actual (vs. predicted for forecasters) own and

partner cooperation rates ($t_s > 4.09, p_s < .001$), caring about partners cooperation ($t = 3.64, p < .001$), and being less tempted to defect ($t = 3.20, p = .001$). Tests of moderation on trust variables confirmed the stakes effects were greater for experiencers than forecasters ($t_s > 6.00, p_s < .001$). Similarly, the stakes effects on liking and relationship-related variables were also greater for experiencers than forecasters ($t_s > 3.25, p_s < .002$), (see Tables 5 and 6, and Figure 1).

Follow-up tests of stakes effects among forecasters (Studies 4 & 5) found few parallel effects to the experiencers (see Tables 4, 5, and 6). Forecasters in the risky (vs. safe) condition cared about their partner’s choices during the PD game more ($b = 0.50, p = .005, d = 0.28$). Marginally more forecasters in the risky (92%) than safe (86%) condition expected their partner to cooperate in the first round ($\chi^2 = 3.26, p = .071$). More forecasters in the risky (94%) condition than the safe (81%) condition expected a typical undergraduate to cooperate in the first round ($\chi^2 = 4.15, p = .042$)—in contrast, fewer forecasters in the risky (30%) condition than the safe (40%) condition predicted full cooperation in all 15 rounds ($\chi^2 = 4.67, p = .031$).

Figure 1: Forecasted Versus Experienced Stakes Effects on Trust Outcomes



Note. Error bars are 95% confidence intervals. Degrees of freedom for experiencers ranged from 276.13-298.33, and for forecasters from 388 to 419.

Specificity: Did Game Stakes Affect Trust More Than Liking?

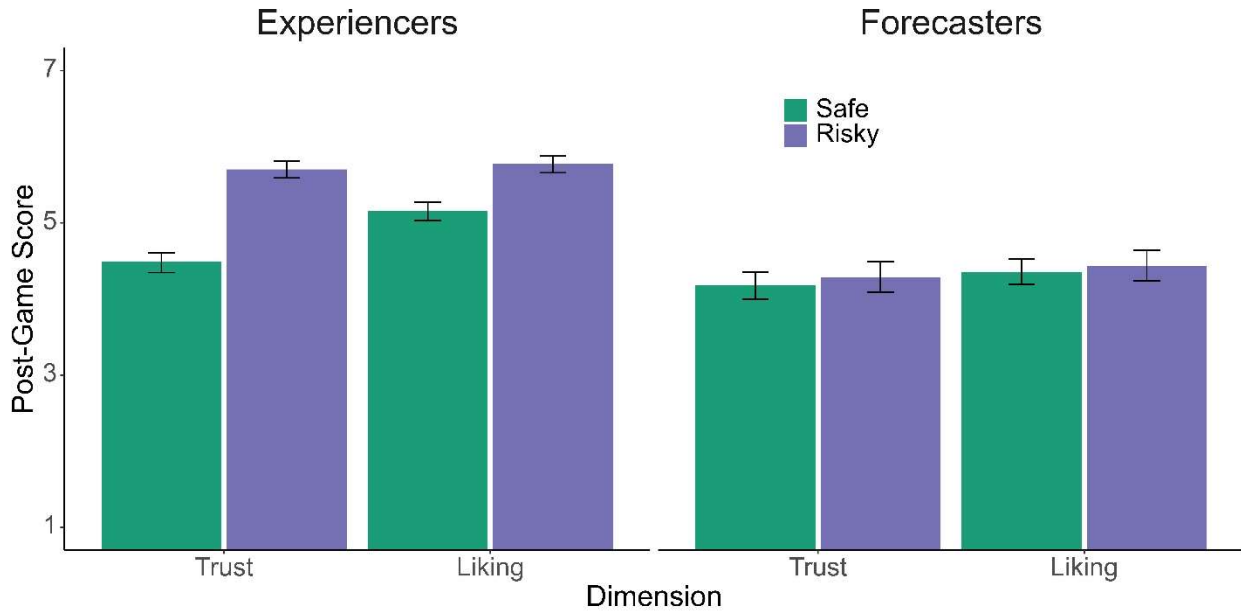
We then compared stakes effects on trust and liking, predicting larger effects on trust than liking because our manipulation was designed to target trust in particular. Inducing high-stakes interdependence in the risky condition was intended to feel trust-diagnostic but not necessarily pleasant or comfortable (because of concerns that one's partner might defect).

Comparing Trust and Liking Directly Among All Participants

We created a difference score, computed as single-item trust minus single-item liking (referred to as “relative trust-over-liking”), to directly test the differential effects of stakes on trust (vs. liking). Using our extended core model, we tested whether relative trust-over-liking varied based on effects-coded stakes, dyad race, and role (including stakes' interactions with dyad race and role). Main effects emerged for both stakes ($b = 0.31, p < .001, d = 0.38$), and role ($b = -0.21, p < .001, d = -0.26$), but these effects were qualified by a stakes-by-role interaction, ($b = 0.58, t = 4.86, p < .001$). Probing this interaction, we find no difference between stakes conditions on the relative trust-over-liking among forecasters ($b = -0.01, t < 1$), whereas experiencers scored higher on relative trust-over-liking in the risky than the safe condition ($b = 0.60, p < .001, d = 0.95$). Given some weak evidence of stakes-by-dyad race interaction ($b = 0.27, t = 1.94, p = .053$), we found that participants in the risky (vs. safe) condition scored higher relative trust-over-liking in same-race dyads ($b = 0.41, p < .001, d = 0.51$), but not cross-race dyads ($b = 0.17, p = .082, d = 0.19$). In all conditions (forecasters or experiencers, safe or risky stakes) the mean of this trust-over-liking score was negative (ranging from -0.07 for experiencers in the risky condition to -0.67 to experiencers in the safe condition), thus participants generally

reported lower trust than liking overall, but the analyses above indicate that the gap was much smaller for experiencers in the risky (vs. safe) condition (see Figure 2).

Figure 2: Post-game Trust and Liking for Experiencers and Forecasters



Note. Error bars are 95% confidence intervals. Descriptive (unadjusted) condition means are reported, but tests of post-game condition differences covaried for respective pre-game measures.

Comparing Experiencer Trust, Liking, and Closeness Gains.

Next, we turned to the slider scores (collected from experiencers only) to extend these comparisons while controlling for pre-game levels of each outcome. We tested effects of PD game stakes on trust, liking, and other relationship factors, with special attention to the comparison of trust and liking, in three steps. First, we used our core model to predict post-game trust, liking, and other relationship factors (separately) while covarying for their respective pre-game levels (see left-hand columns of Table 7). Although stakes affected all outcomes significantly (all $ps < .001$), post-game trust was substantially higher in the risky ($M = 4.84$), than the safe ($M = 3.95$) condition, whereas this gap between the risky ($M = 4.97$), and safe ($M = 4.62$) conditions was smaller for post-game liking (see Table 7). This asymmetry was driven by

the large trust increase from pre-game to post-game in the risky (but not safe) condition ($M_{\text{increase}} = 0.78$), and a slight decrease from pre-game to post-game liking in the safe (but not risky) condition ($M_{\text{decrease}} = 0.25$). To test if trust gains were solely due to induced closeness from the pre-game closeness tasks, we analyzed the stakes effects on trust with pre-game closeness as a factor interacting with stakes, in which stakes remained the strongest predictor of trust, and no moderation of stakes' effects on trust by pre-game closeness was found (for details, see Appendix L).

Table 7: Stakes Effects on Experiencers' Post-Game (vs. Pre-Game) Outcomes

	Post-game measure		Change (Post – Pre)					
	Safe	Risky	Stakes Effect	Safe	Risky	Stakes Effect		
Slider measures	$M (SD)$	$M (SD)$	t	d	$M (SD)$	$M (SD)$	t	d
Trust	3.95 (1.36)	4.84 (1.38)	10.56 ^{***}	1.24	0.04 (0.80)	0.78 (0.98)	9.71 ^{***}	1.14
Liking	4.62 (1.18)	4.97 (1.22)	6.09 ^{***}	0.72	-0.25 (0.64)	0.07 (0.64)	6.05 ^{***}	0.71
Closeness	2.95 (1.18)	3.69 (1.33)	9.28 ^{***}	1.09	0.14 (0.74)	0.71 (0.75)	8.89 ^{***}	1.04
IOS	3.40 (1.36)	4.02 (1.38)	9.34 ^{***}	0.60	0.05 (0.66)	0.52 (0.63)	8.85 ^{***}	1.03
Similarity	3.83 (1.35)	4.29 (1.42)	5.99 ^{***}	0.72	-0.06 (0.86)	0.36 (0.98)	5.56 ^{***}	0.66
Interest	4.56 (1.35)	4.99 (1.37)	5.79 ^{***}	0.71	-0.19 (0.75)	0.16 (0.81)	5.53 ^{***}	0.69

Note. Boldface indicates significant effects. IOS = inclusion of other in self. N for each variable = 578. All measures were scored from 1-7. Descriptive (unadjusted) condition means are reported, but tests of post-game condition differences covaried for respective pre-game measures.

*** $p < .001$. ** $p < .01$. * $p < .05$. † $p < .1$.

Next, we calculated change scores (i.e., subtracted pre-game trust from post-game trust, and the same for liking), which we submitted to our core model to assess the differing impact of stakes on trust versus liking for experiencers (see right-hand columns of Table 7). These results aligned closely with those in the previous step, in terms of magnitude and significance.

Finally, in a long-form version of our dataset (with pre- and post-game trust and liking slider scores in separate rows for each participant) we added a dimension (trust vs. liking) variable to the core model to test directly whether the stakes effect on post-game trust versus liking varies. Analyses covaried for the pre-game score for the respective dimension, as

measured on the same sliders. We found the predicted interaction of stakes and dimension ($b = 0.44, t = 5.23, p < .001$), confirming that PD game stakes affected trust more so than liking, when controlling for pre-game levels of each.

Extending Gains to a Novel Situation: Negotiation Effects (Study 2-3)

We then tested whether the stakes effects extend to the negotiation task in Studies 2 and 3. We predicted that the experiencers in the risky (vs. safe) condition would report more trust, satisfaction and other positive outcomes during and after the negotiation, and that post-game trust would mediate these positive negotiation outcomes.

Stakes Main Effects on Negotiation Outcomes

Stakes effects for experiencers emerged on a few post-negotiation outcomes (See Table 8). Experiencers in the risky (vs. safe) condition reported more trust during the subsequent negotiation ($b = 0.41, p = .003, d = 0.42$) and marginally better treatment from their partner ($b = 0.24, p = .074, d = 0.25$). Experiencers in the risky (vs. safe) condition also reported a marginal increase of open sharing of information between partners ($b = 0.35, p = .081, d = 0.25$) and significantly greater understanding between themselves and their partner ($b = 0.25, p = .049, d = 0.28$). Finally, from the perceived change items, experiencers reported multiple marginal effects: increased closeness ($b = 0.22, p = .079, d = 0.27$), a stronger relationship ($b = 0.21, p = .049, d = 0.25$), and a better relationship ($b = 0.23, p = .073, d = 0.25$), with their partner after the negotiation task.

Table 8: Main Effects of Stakes on Negotiation Outcomes

Negotiation Outcomes	Safe	Risky	Stakes Main Effect	
	Mean (<i>SD</i>)	Mean (<i>SD</i>)	<i>t</i>	<i>d</i>
Trust	5.13 (1.30)	5.56 (1.10)	2.96**	0.42
Trust (perceived change)	4.85 (1.22)	5.00 (1.16)	1.14	0.16
Satisfaction	5.45 (1.34)	5.52 (1.16)	0.57	0.08
Partner treatment	5.65 (1.28)	5.90 (1.20)	1.79 [†]	0.25
Relationship satisfaction	5.61 (1.19)	5.74 (1.15)	0.93	0.13
Engagement	5.27 (1.26)	5.28 (1.21)	0.12	0.02
Info sharing	4.81 (1.58)	5.12 (1.62)	1.75 [†]	0.25
Comfort	5.30 (1.22)	5.39 (1.17)	0.65	0.09
Understanding	5.20 (1.11)	5.44 (1.14)	1.98*	0.28
Partner warmth and competence	5.41 (1.06)	5.59 (1.02)	1.50	0.21
Cultural differences	3.00 (1.63)	2.90 (1.61)	0.78	-0.11
More pleasant (perceived change)	4.84 (1.21)	5.02 (1.16)	1.36	0.19
Closer (perceived change)	4.58 (1.05)	4.81 (1.09)	1.90 [†]	0.27
Stronger (perceived change)	4.61 (1.08)	4.82 (1.11)	1.77 [†]	0.25
Better (perceived change)	4.79 (1.16)	5.02 (1.12)	1.80 [†]	0.25

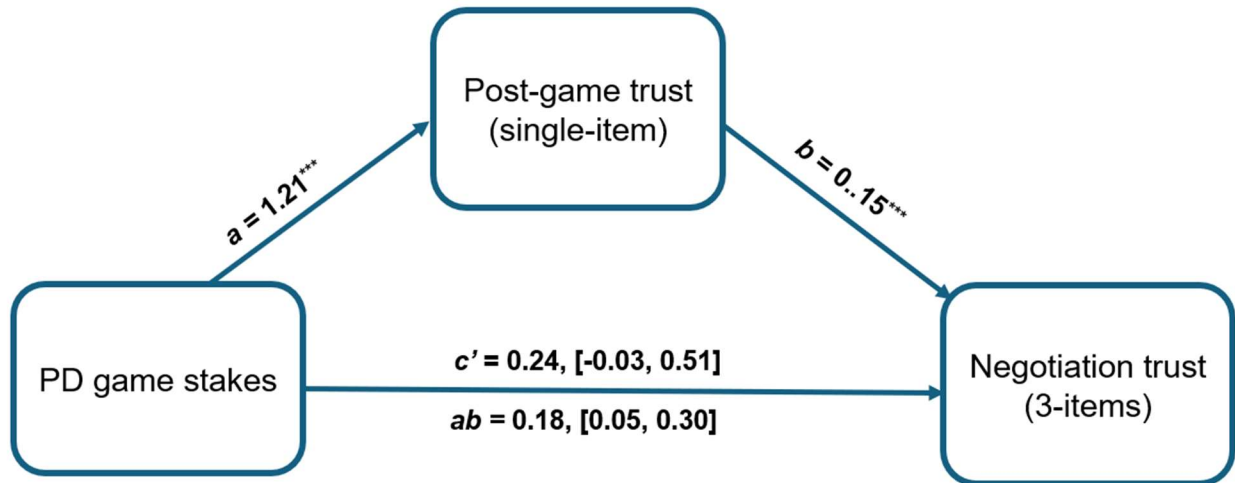
Note. Boldface indicates significant effects. *N* for all variables = 395.

*** $p < .001$. ** $p < .01$. * $p < .05$. [†] $p < .1$.

Multi-level Mediation of Stakes Predicting PD Game Trust Predicting Negotiation Trust

To test that increased trust from the PD game specifically is associated with increased trust in the negotiation we used a multi-level mediation model, specified with post-game trust mediating the effect of risky (vs. safe) on negotiation trust (for all path values, see Figure 3). The indirect effect of stakes through post-game single-item trust was significant, ($b = 0.18$, 95% $CI = [0.05, 0.30]$). After accounting for mediation, the direct effect of stakes on negotiation trust dropped to non-significance, ($b = 0.24$, 95% $CI = [-0.03, 0.51]$).

Figure 3: Stakes Effect on Negotiation Trust Mediated by PD Game Trust



Note. Post-game trust and negotiation trust were scored from 1-7, negotiation trust was a composite of three items. ab = indirect effect, c' = direct effect.

*** $p < .001$. ** $p < .01$. * $p < .05$. † $p < .1$

Does Cooperation (Actual or Predicted) Mediate Effects of Stakes on Trust?

Finally, we investigated cooperation as a potential mechanism for the stakes effects. We predicted that the cooperation rate within a dyad would be a key mechanism by which stakes impact post-game trust, and that forecasters would fail to appreciate the relationship between stakes, cooperation, and trust.

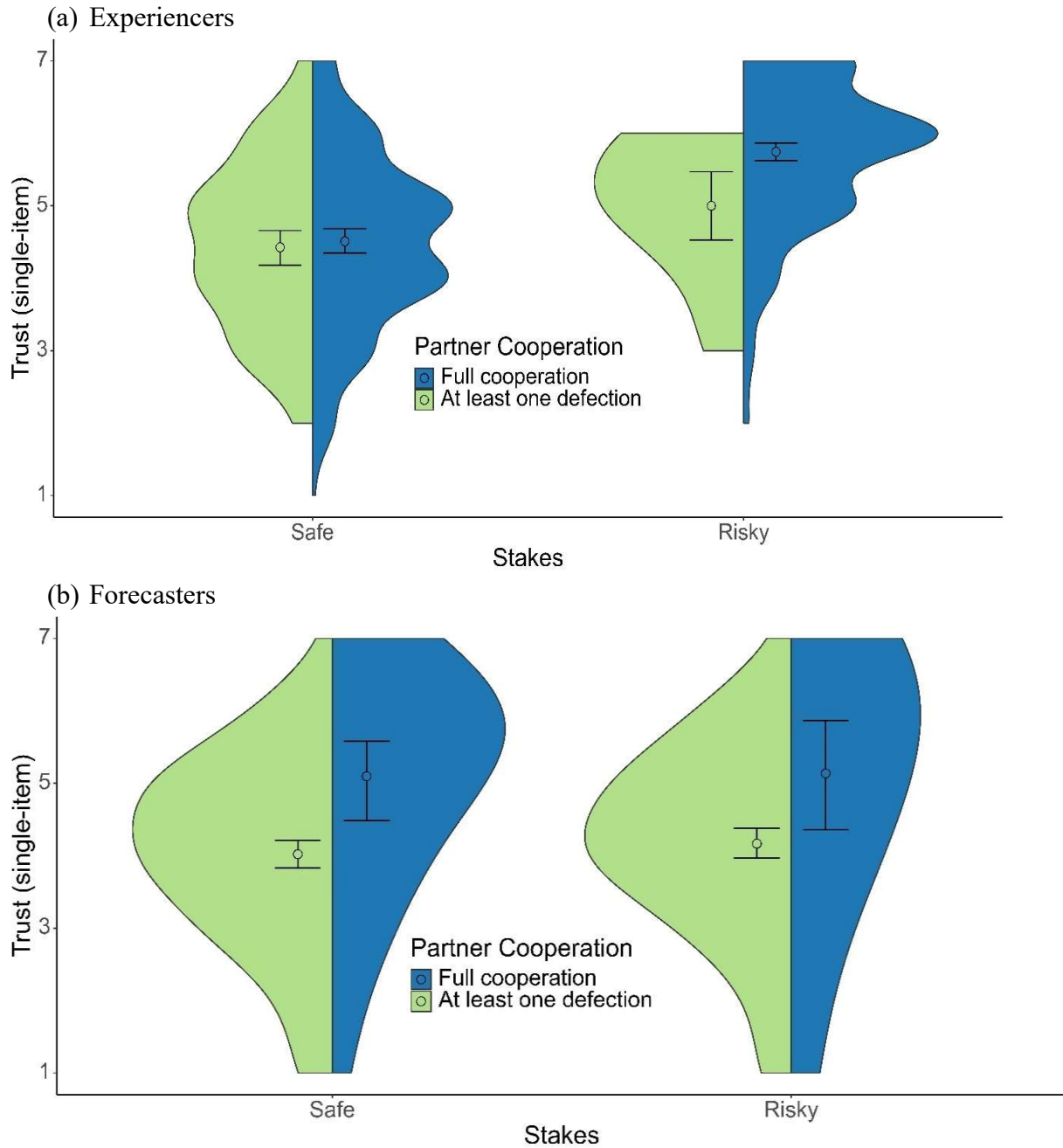
First, we illustrate the importance of cooperation by examining trust distribution when one's partner cooperated for all 15 rounds versus if one's partner defected even once. Then we formally test cooperation as the mechanism by which stakes affects trust using a multi-level moderated mediation with parallel cooperation mediators.

To illustrate the damage of betrayal (defection in the PD game) on subsequent trust building, we conducted additional analyses with full (actual or predicted) partner cooperation. This variable was entered into regressions predicting single-item trust with the effects-coded stakes and dyad race variables and their interactions for experiencers and forecasters, separately. This model additionally covaried for full cooperation from the participant (or predicted full

cooperation for forecasters) The partner's complete cooperation (vs. defecting at least once), had a significant main effect ($b = 0.36, p = .024$), and a marginal interaction with stakes ($b = 0.60, t = 1.90, p = .058$). Cautiously probing this marginal interaction, we see that in the safe condition partner's complete cooperation (vs. defecting at least once) has essentially zero effect on trust ($b = 0.06, p = .696$), which is expected as participants were not aware of partner cooperation in the safe condition. In the risky condition, however, partner's full cooperation (vs. defecting at least once) predicted trust ($b = 0.66, p = .020$). Forecasters predicted only a significant main effect of partner's complete cooperation ($b = 0.85, p < .001$).

Visually inspecting the data in the safe and risky conditions when one's partner in the dyad always cooperated (vs. defected at least once) reveals marked differences in their distributions (see Figure 4). In the risky condition, if the participant's partner defected at least once, the participant never reported complete trust in their partner after the PD game, while in the safe condition both distributions are quite similar. The forecaster data distribution shows that forecasters do understand the impact of cooperation. Across risky and safe conditions they predict higher trust when they predict their partner will never defect, however they fail to appreciate the stakes effect we see in the experiencers—the risky and safe distributions are virtually identical using their predictions.

Figure 4: Experiencers' and Forecasters Post-Game Trust by Partner's Full Cooperation



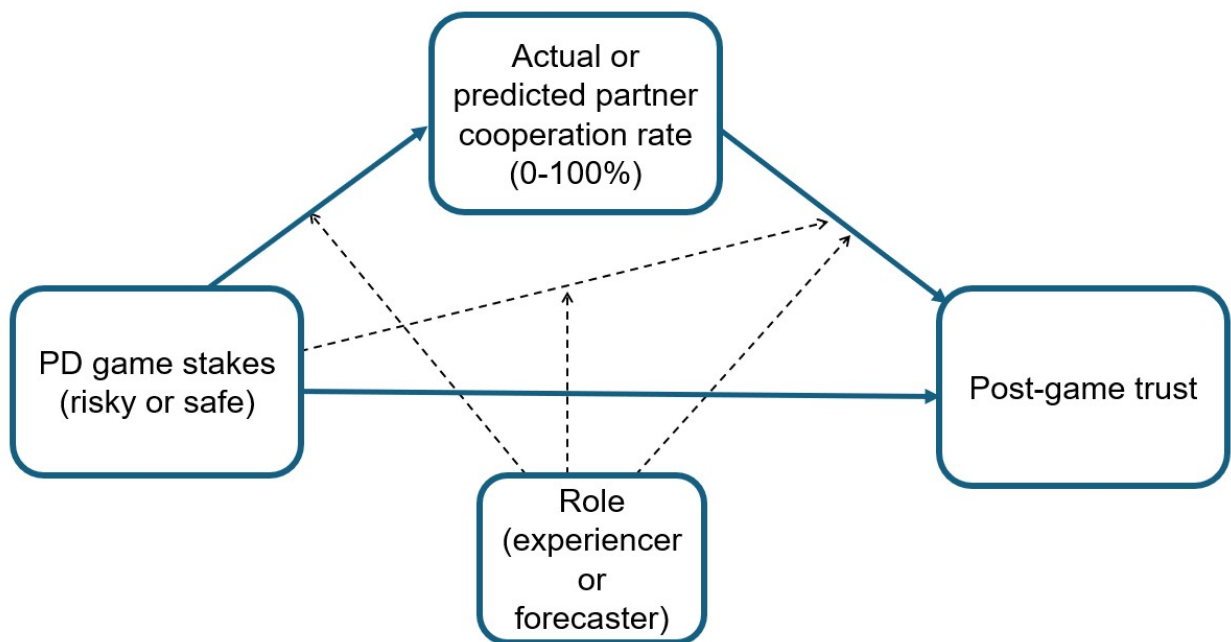
Note. Error bars are 95% confidence intervals. Panel (a) shows the distribution of data for experiencers in each stakes condition, split by whether their partner cooperated for all 15 rounds (vs. defected at least once). Panel (b) shows the same distributions for predicted cooperation in forecasters.

Moderated Mediation of Stakes Predicting Trust Through Cooperation

Next, we ran regression and multilevel moderated mediation analyses to test the core hypotheses that motivated this line of research, which was designed to test whether a stakes

manipulation designed to elicit higher rates of cooperation could effectively induce trust, as well as whether such effects are non-obvious to lay people (here, forecasters). Our basic hypotheses involving mediation were that risky (vs. safe) stakes in the PD game would lead to a higher rate of cooperation (by both self and partner) and then, in turn, that partner’s cooperation—specifically for those in the risky condition—would increase trust. Thus, this model includes an interaction between the stakes manipulation and the mediator, as tested in contemporary causal mediation models. We further extended this model by adding role as a moderator of the *a* and *b* paths, as well as the interaction between stakes and the proposed mediator (partner cooperation), to test whether processes unfolded differently for experiencers versus forecasters (see Figure 5).

Figure 5: Theoretical Stakes by Cooperation by Role Mediation on Trust



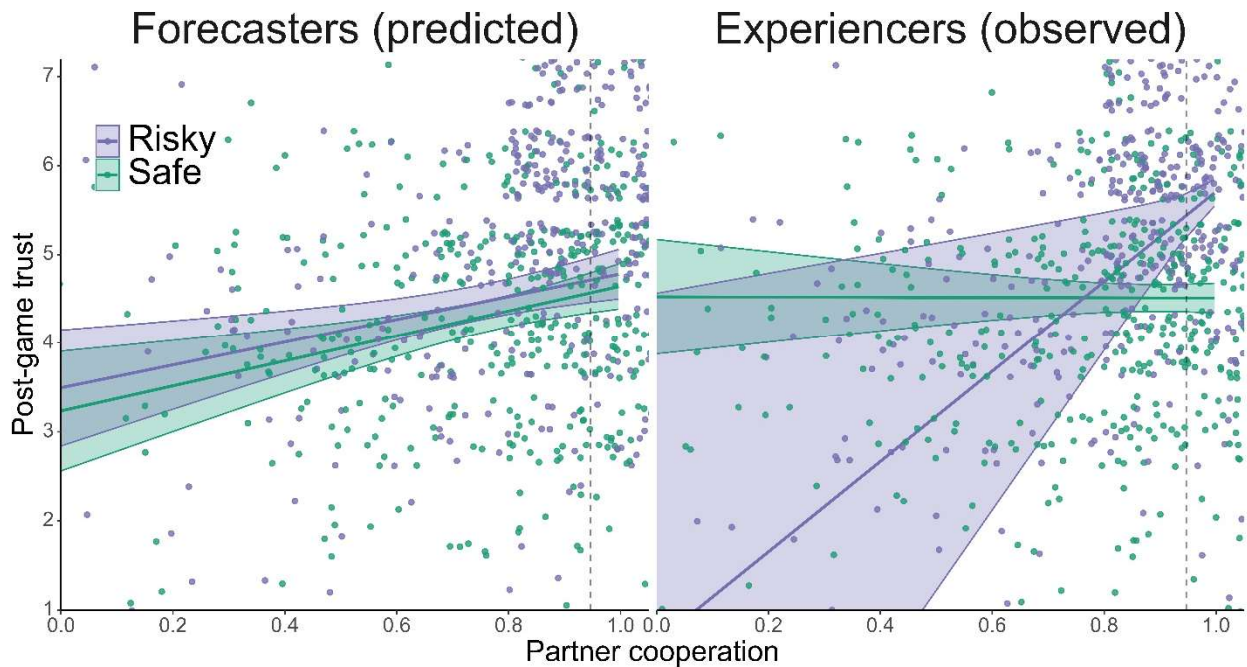
Note. Dashed arrows indicate interaction terms.

First, to test whether the effects of PD game stakes and partners’ cooperation during the PD game on post-game trust varied for experiencers versus forecasters, we extended our core model (including dyad race) to regress the single-item post-game trust measure on the three-way interaction of role, stakes, and partner cooperation (actual for experiencers, predicted for

forecasters), while covarying for own (actual or predicted) cooperation. Cooperation rate was mean-centered based on actual (i.e., experiencer) cooperation. As hypothesized, both one's own cooperation ($b = 0.54, p = .038$) and one's partner's cooperation ($b = 1.94, p < .001$) during the PD game positively predicted trust levels after the PD game, although the unique effect of partner (vs. own) cooperation was descriptively larger.

Turning to our questions about moderation, as expected, the association between cooperation and trust was moderated by stakes (Cooperation \times Stakes: $b = 2.48, t = 2.36, p = .019$). Notably, although the strength of the overall association between actual or predicted cooperation and post-game trust did not vary for experiencers versus forecasters (Cooperation \times Role: $b = 1.18, t = 1.10, p = .272$), this null effect was qualified by the predicted three-way interaction involving stakes (Cooperation \times Role \times Stakes: $b = 5.20, t = 2.47, p = .013$). We probed this interaction separately for each role. We found a cooperation-by-stakes interaction among experiencers ($b = 4.78, t = 2.57, p = .010$), but not forecasters ($b = -0.16, t = -0.30, p = .766$; see left panel, Figure 6), indicating that forecasters anticipated that their partners' cooperation would affect their trust comparably regardless of PD game stakes. As expected, among experiencers, partners' cooperation predicted experiencers' post-game trust in the risky condition ($b = 4.79, p = .009$), but not the safe condition ($b = 0.01, p = .977$), in which information about partners' cooperation was kept secret (see right panel, Figure 6).

Figure 6: Cooperation by Stakes on Post-game Trust, Split by Role



Note. Shaded area indicates 95% confidence interval. Values covary for own (predicted or observed) cooperation rate, at the mean of observed cooperation (represented by the dashed vertical line).

Building on these results, we then estimated multi-level mediation models separately for experiencers and forecasters to quantify the relevant indirect effects and assess whether this mediation pattern generalized to our other operationalizations of trust. All mediation models tested whether effects of stakes on trust were mediated by the partner's rate of cooperation. The b path was specified with stakes interacting with partner's rate of cooperation, and covaried for the participants' own rate of cooperation. This approach enables us to estimate the magnitude of the natural indirect effect (Montoya, 2024) specific to each stakes condition. All mediation models additionally covaried for dyad race and its interaction with stakes, for equivalence to our core model. The natural indirect and direct effects were calculated using 1000 bootstrap simulations. The mediations revealed extremely similar patterns across trust operationalizations (see Table 9).

Table 9: Mediation Path Coefficients for Trust and Relationship Outcomes

Trust outcomes	Risky		Safe		Stakes × Coop <i>b</i> (<i>SE</i>)	<i>c'</i> path <i>b</i> [95% <i>CI</i>]
	<i>b</i> path <i>b</i> (<i>SE</i>)	<i>ab</i> path <i>b</i> [95% <i>CI</i>]	<i>b</i> path <i>b</i> (<i>SE</i>)	<i>ab</i> path <i>b</i> [95% <i>CI</i>]		
Experiencers						
Trust (single-item)	4.79** (1.82)	0.46 [0.12, 0.82]	0.01 (0.32)	0.002 [-0.06, 0.06]	4.78** (1.85)	0.85 [0.58, 1.11]
Reliance (single-item)	4.30* (2.05)	0.41 [0.03, 0.82]	0.46 (0.37)	0.04 [-0.03, 0.12]	3.84 [†] (2.08)	0.74 [0.45, 1.03]
Trust (PD game impact)	3.47 (2.70)	0.34 [-0.15, 0.86]	0.70 (0.48)	-0.06 [-0.16, 0.03]	4.17 (2.74)	1.75 [1.37, 2.13]
Trust (perceived change)	4.61** (1.60)	0.44 [0.14, 0.77]	0.42 (0.29)	0.04 [-0.02, 0.10]	4.19* (1.63)	1.04 [0.8, 1.26]
Trust (post-game slider)	5.82*** (1.38)	0.54 [0.28, 0.87]	0.08 (0.25)	-0.01 [-0.05, 0.04]	5.91*** (1.40)	0.44 [0.21, 0.66]
Forecasters						
Trust (single-item)	2.09*** (0.53)	-0.07 [-0.19, 0.04]	2.26*** (0.55)	-0.07 [-0.2, 0.04]	0.16 (0.55)	0.17 [-0.09, 0.45]
Reliance (single-item)	2.64*** (0.56)	-0.08 [-0.23, 0.05]	2.01*** (0.58)	-0.06 [-0.18, 0.04]	0.63 (0.58)	0.03 [-0.25, 0.32]
Trust (PD game impact)	1.78** (0.64)	-0.06 [-0.18, 0.03]	0.66 (0.65)	-0.02 [-0.1, 0.03]	1.12 [†] (0.66)	0.12 [-0.2, 0.45]
Trust (perceived change)	1.87*** (0.46)	-0.06 [-0.17, 0.03]	1.36** (0.49)	-0.04 [-0.13, 0.02]	0.51 (0.49)	0.19 [-0.04, 0.44]

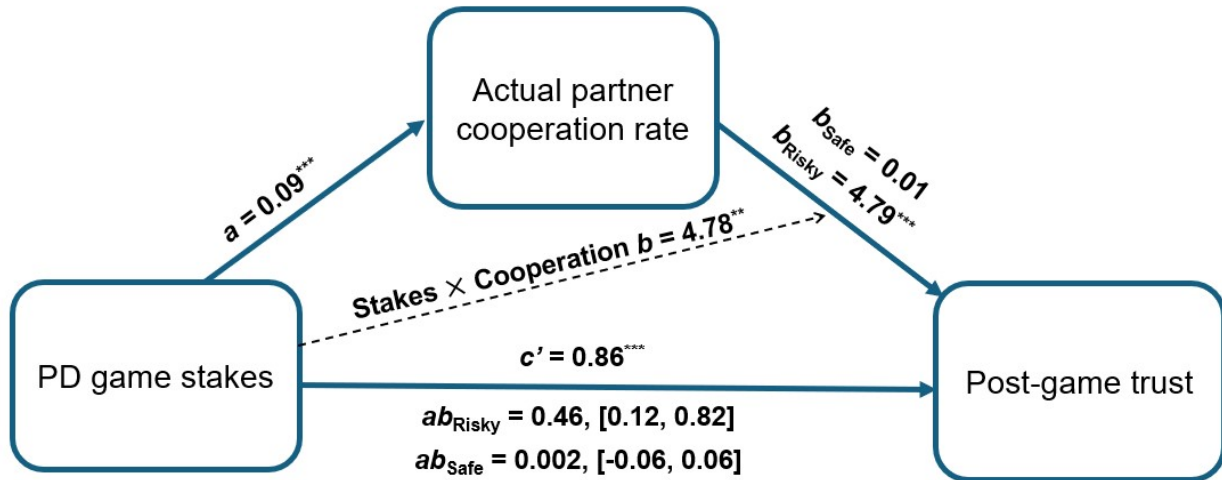
Note. Boldface indicates significant effects. Experiencer unstandardized *a* path = 0.09, forecaster *a* path = -0.03. *b* paths are simple effects of partner cooperation within each stakes condition. *ab* = natural indirect effect, *c'* = natural (controlled) direct effect. All *b* paths, indirect effects, and direct effects covaried for own cooperation.

*** $p < .001$. ** $p < .01$. * $p < .05$. [†] $p < .1$

Among experiencers, four of the five trust-related outcomes showed the following pattern, which we report in text only for post-game single-item trust (corresponding to Figure 7). Risky (vs. safe) PD game stakes led to greater partner cooperation (*a* path = 0.09, $p < .001$). In turn, partner cooperation predicted post-game trust in the risky condition (b_{Risky} path = 4.79, $p < .001$), but not the safe condition (b_{Safe} path = 0.01, $p = .979$) and these paths differed significantly (Stakes × Partner cooperation interaction: $b = 4.78$, $p < .001$). The indirect effect from stakes via partner cooperation to trust was significant in the risky condition ($ab_{\text{Risky}} = 0.46$, 95% CI from 0.12 to 0.82) but not the safe condition ($ab_{\text{Safe}} = 0.002$, 95% CI from -0.06 to 0.06).

Finally, the direct effect of stakes on trust persisted after accounting for partner (and own) cooperation (c' path = 0.86, $p < .001$).

Figure 7: Stakes Predicting Trust Mediated by Partner Cooperation for Experiencers



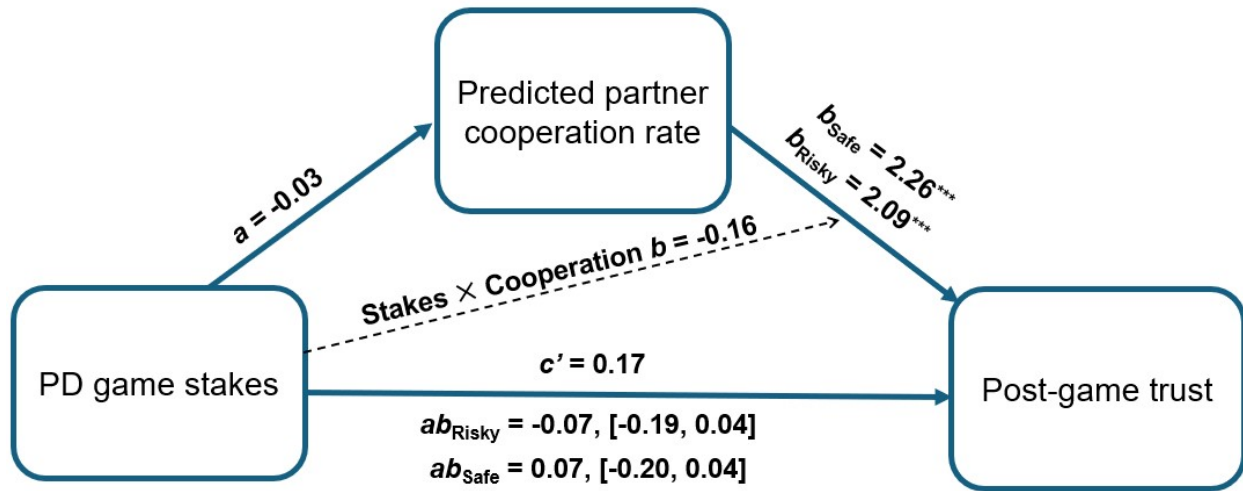
Note. Cooperation rate was measured by averaging participants' cooperation across all 15 rounds, single-item trust was measured using a 1-7 scale. Dashed line indicates the interaction term. All terms except the a path are estimated with the participant's own cooperation rate as a covariate to isolate the unique impact of partner cooperation rate. ab = natural indirect effect, c' = natural (controlled) direct effect.

*** $p < .001$. ** $p < .01$. * $p < .05$. † $p < .1$

Among forecasters, all four trust-related outcomes showed the following pattern, which we report in text only for post-game single-item trust (corresponding to Figure 8). Risky (vs. safe) PD game stakes did not lead to greater predicted partner cooperation (a path = -0.03, $p = .196$). However, predicted partner cooperation did predict post-game trust in the risky condition (b_{Risky} path = 2.09, $p < .001$), and the safe condition (b_{Safe} path = 2.26, $p < .001$) and these paths were not moderated by stakes (Stakes \times Partner cooperation interaction: $b = -0.16$, $p = .766$). The indirect effect from stakes via predicted partner cooperation to trust was not significant in the risky condition ($ab_{\text{Risky}} = -0.07$, 95% CI from -0.19 to 0.04) or the safe condition ($ab_{\text{Safe}} = -0.07$, 95% CI from -0.20 to 0.04). Finally, the direct effect of stakes on trust

remained nonsignificant after accounting for predicted partner (and own) cooperation (c' path = 0.17, $p = .212$).

Figure 8: Stakes Predicting Trust Mediated by Partner Cooperation



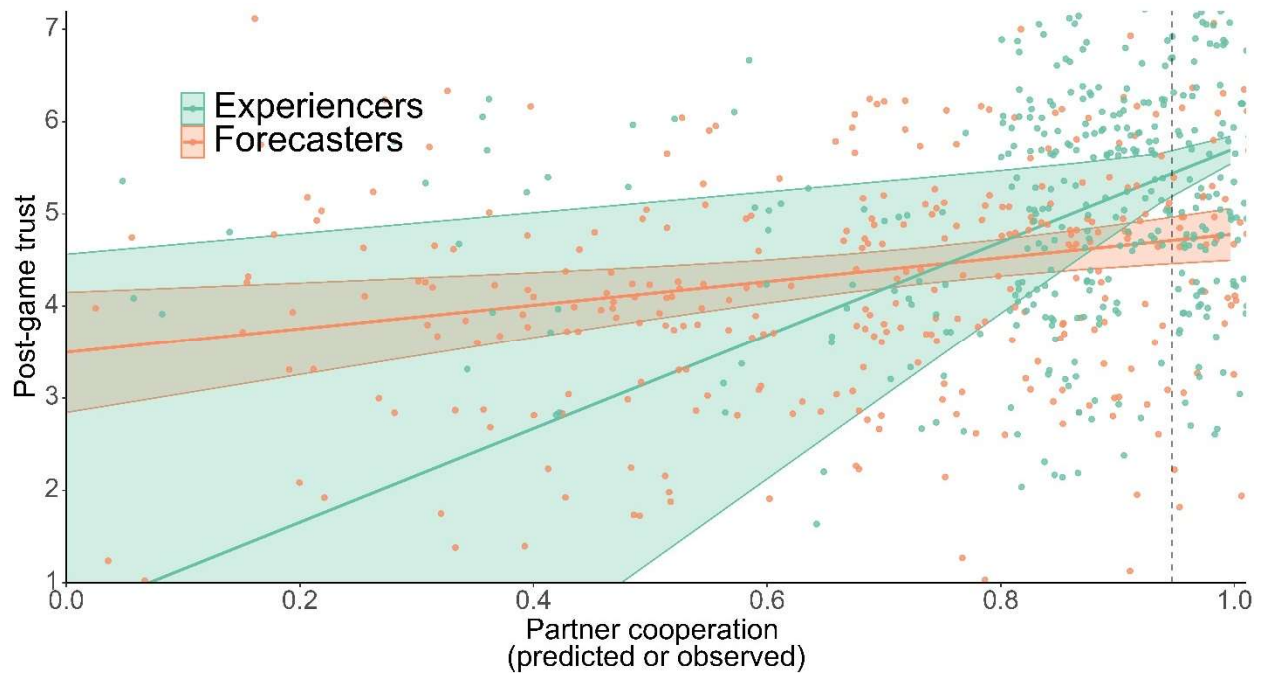
Note. Partner cooperation rate was predicted by forecasters using a seven-point scale, which was then scaled from 0 (never cooperate) to 1 (always cooperate) for comparisons to experiencers. Post-game trust was measured using a 1-7 scale. Dashed line indicates the interaction term. All terms except the a path are estimated with the participant's own cooperation rate as a covariate to isolate the unique impact of partner cooperation rate. ab = natural indirect effect, c' = natural (controlled) direct effect.

*** $p < .001$. ** $p < .01$. * $p < .05$. † $p < .1$

Finally, we return to the three-way interaction of cooperation, role and stakes, to shed some light on what exactly the forecasters are missing when predicting PD game effects. In addition to the two- and three-way interactions there was a significant main effect of role across risky and safe conditions (at the mean of actual cooperation), such that experiencers (vs. forecasters) reported higher trust ($b = .33, p = .003$). We can also break down the three-way interaction by looking at the simple interactions of role and cooperation within the risky (vs. safe) condition (rather than looking at the simple interactions of stakes and cooperation within each role, as was done earlier). This allows us to see the pattern of results more clearly within each stakes condition (compared to each role). In the safe condition, the association between

(actual) partner cooperation and trust is significantly weaker among experiencers (who don't have access to the cooperation info) than the predicted relationship among forecasters, $b = -1.42$, $p = .011$. However, in the risky condition this pattern reverses, such that the association between (actual) partner cooperation and trust is marginally stronger among experiencers than the predicted relationship among forecasters, $b = 3.78$, $p = .066$ (see Figure 9).

Figure 9: Cooperation by Role on Post-game Trust in the Risky Condition



Note. Shaded area indicates 95% confidence interval. Values covary for own (predicted or observed) cooperation rate, at the mean of observed cooperation. The dashed line indicates the mean of observed cooperation.

CHAPTER 4: DISCUSSION

Optimal functioning, both in work and personal relationships, requires trust.

Understanding how to build trust and reap its benefits has immediate real-world applications (e.g., improving wellbeing & mental health; Zhao et al., 2024). The present research tested the effectiveness of a novel PD game intervention designed to induce interdependence, and thereby trust, yielding interpersonal benefits that can carry forward into new contexts (Kelley & Thibaut, 1978; Simpson, 2007). Five studies demonstrate the effectiveness of this intervention at improving cooperation and subsequently trust for people who actually experience it, as well as its subtlety, insofar as the resulting trust gains are not obvious without experiencing the intervention firsthand: People who imagine this full paradigm fail to accurately forecast the resulting trust gains. Below we discuss several remaining questions, limitations, and future directions.

Persisting Trust Gained Following Cooperation

The risky, high-stakes PD game had robust impacts on trust and cooperation rates. We also found support for partner cooperation as a (partial) mechanism by which this high-stakes PD game improves trust. We used a study design that enabled us to test a “gold standard” model that reflects best practices for mediations: temporal separation between the independent variable (stakes), hypothesized mediator (partner cooperation rate), and outcome (trust) variables, each implemented or assessed in the hypothesized causal sequence (see Holland et al., 2017); experimentally manipulating the independent variable *and* (participants’ access to) the hypothesized mediator (Spencer et al., 2005); and formally testing moderation by stakes of the effect of partner cooperation on trust (i.e., testing the treatment-mediator interaction; see Rijnhart et al., 2021). In other words, because only the risky condition enabled participants to know of

their partners' cooperation rate (whereas this information was kept secret in the safe condition), our manipulation of stakes additionally served as a pseudo-manipulation of the mediator. Thus, our finding that partner cooperation accounts for a substantial portion of stakes' effect on trust avoids the usual concerns associated with testing mediation using observational cross-sectional data (e.g., see Fiedler et al., 2018).

The PD game was designed to promote cooperation (and was successful, as evidenced by the extremely high rates of cooperation in the high-stakes PD game), because cooperation in the PD game signals trustworthiness to their partner. Not unlike strain-test situations in close relationships, putting someone else's needs above one's own is an especially powerful mechanism to build trust when one has the option to instead act selfishly, for example in risky interdependence situations like the high-stakes PD game (Holmes, 1981). The fundamental attribution error also likely factors into these results; the observed effects on trust could partially be due to participants' tendency to over-attribute dispositional motivations for others' behaviour, rather than situational factors (Ross, 1977). Thus, participants are discounting the powerful situation they (and their partner) are in, instead inferring that their partner is especially trustworthy due to their high cooperation; however, cooperating highly was overwhelmingly common by design.

Trust gains from this brief PD game also persisted into a novel interactive context, suggesting that these trust gains may withstand at least some subsequent stress-testing. To our knowledge, no other study has extended trust gains from one social dilemma into another type of task (in our case, the negotiation) with dyads, though some work has explored such effects with individuals and computer agents (Collins et al., 2016). Though beyond the scope of the present research, investigating how these gains generalize beyond sequential in-lab tasks is a promising

area for future research with clear real-world implications for building trusting relationships in personal and professional contexts. Few studies have investigated longitudinal impacts on trust from social dilemmas, or their generalizability to other types of trust (e.g., trusting someone to care for your emotions), or other contexts. Determining how far we can push these effects remains a beneficial line of further inquiry, clarifying its immediate practical implications for team-building, relationship repair, and other applied settings.

Forecasters Underestimate Cooperation Rates

Forecasters underestimated overall cooperation rates and completely failed to anticipate the impact of the high-stakes PD game versus the low-stakes PD game on cooperation rates. This underestimation likely reflects a lack of appreciation for how social norms promoting pro-social behaviour, combined with the binary cooperate/defect PD game format, can constrain people's choices. People want to be seen as a good person, and the minimal ambiguity in PD games increases the difficulty in protecting yourself from negative interpretations of your actions—few alternative, face-saving explanations can account for a defection, and defections create lasting negative impacts (Lount et al., 2008). Moreover, our high-stakes PD game enables partners to see your choices, making very salient that your partner could (and perhaps should) judge you based on your actions. Additionally, a recent meta-analysis found that communication between partners in social dilemmas, even when it took place only before the actual dilemma, was one of the strongest predictors of cooperation (Jin et al., 2025). Thus, forecasters may also have underestimated the benefits of the closeness-inducing tasks (not only inducing closeness but allowing communication between partners), and the downstream effects that it might have on cooperation. Forecasters, not being in the situation themselves, were likely not aware of all the

ways that the high-stakes interdependence-inducing PD game influences one's behaviour, or how powerful the effects could be.

Although forecasters did correctly predict that greater partner cooperation would increase trust, they failed to appreciate the impact of PD game stakes on either cooperation or trust (and on the association between the two). Our final moderation model found that forecasters still underestimated post-game trust (relative to actual levels reported by experiencers), even among those who accurately forecasted (high) levels of partner cooperation (i.e., when probing role simple effects at the mean of actual partner cooperation). Future studies could test whether this underestimation replicates even when experimentally manipulating the cooperation-related information given to forecasters enables them to anticipate the trust gains observed in the high-stakes condition for experiencers. Despite our moderation model finding trust underestimation even when predicted cooperation matched actual cooperation, forecasters might be able to correctly estimate post-game trust levels if told that in the high-stakes PD game their partner is 94% likely to cooperate in all 15 rounds (i.e., full cooperation). However, the previously cited meta-analysis of cooperation games found an average rate of cooperation across 2340 studies just below 50% (Jin et al., 2025), plus experiencers in our studies who were in this very situation expected typical undergraduates to cooperate at lower levels (77%), so convincing the forecasters of the very high cooperation rate in the high-stakes PD game could prove difficult.

Interestingly, another study using the PD game found that participants were able to predict partner cooperation at above-chance levels after having interacted with them, but the participants' error stemmed from *overestimating*, not underestimating, partner cooperation, opposite to comparing forecasters with experiencers in the present study (Sparks et al., 2016). In their publicly available data both actual (74%) and predicted (82%) cooperation rates fell well

below our experiencers' actual rate of cooperation (95%), although their actual rate of cooperation resembled our forecasters' predicted rate (70%). This convergence suggests that perhaps our forecasters are predicting cooperation more in line with a typical PD game, and that our high-stakes 15-round PD game is what is causing forecasters to mispredict cooperation.

Notably, the present study was designed and tested in a relatively high-trust culture (a large public university in Canada), and so the social norm is to trust (Falk et al., 2018). However, in other cultures or high-conflict settings trusting (especially strangers) may be less normative, and in these contexts participants may be less likely to cooperate, causing the PD game to backfire. Thus, the predictions from forecasters on cooperation rates and subsequent (lack of) trust building might also be more accurate in low-trust societies (e.g., South Africa, Falk et al., 2018).

Specific Gains on Trust Versus Liking

People often tend to start out at a relatively high level of liking others (as evident in the person positivity effect; Sears, 1983), whereas trust needs to be earned and can be difficult to build. Indeed, some theoretical models of trust development posit that “trust begins at zero when no prior information is available” (Lewicki et al., 2006, p. 994). Given general defaults to like more than trust others, the differential impact of the high-stakes PD game on trust, over and above liking, is especially impressive. Our pattern of results aligns with the literature on relationships and interdependence (Kelley & Thibaut, 1978; Sears, 1983). Liking starts high, trust starts comparatively lower, and inducing interdependence improves specifically trust, eliminating the pre-game trust-liking gap after completing only the high-stakes PD game (as no such trust increase occurred in the low-stakes PD game).

Closeness is another piece of the puzzle—the closeness-inducing tasks were included for all participants to help encourage cooperative behaviour; however, this design feature makes it difficult to disentangle the stakes effects from the closeness-inducing tasks. A certain level of closeness may be necessary to reap the benefits of the risky PD game, with closeness-inducing tasks playing a role in supporting sufficiently high cooperation rates for trust to emerge. This account, however, seems somewhat at odds with the pattern of data observed here. After the closeness-inducing tasks (and prior to the PD game), experiencers' average closeness scores were below the scale midpoint—major gains in trust, closeness, or liking were evident only after the risky (not safe) PD game. Moreover, pre-game closeness did not moderate the effect of PD game stakes on post-game trust (see appendix L), suggesting that the closeness induction likely did not play a major role in producing the observed trust gains.

Therefore, although the present research cannot directly address the question of whether the closeness-inducing tasks are necessary for the high-stakes condition to effectively build trust, the stakes manipulation strongly impacts trust regardless of participants' level of closeness they feel with their partner. Moreover, including the closeness-inducing getting-acquainted tasks—rather than testing effects of our PD game paradigm at literally zero-acquaintance—increases the generalizability of our results to the real-world personal and work relationships characterized by some level of pre-existing interpersonal familiarity or closeness.

Similar Effects for Same- and Cross-race Dyads

That dyad race did not moderate effects of PD game stakes on our focal outcomes suggests that the high-stakes risky (vs. low-stakes safe) PD game was comparably effective inducing interdependence (and cooperation, followed by trust) for both same- and cross-race dyads. Given multiple theoretical reasons to expect some form of moderation by racial

composition (e.g., same-race dyads may be more likely to cooperate due to sharing an identity, or cross-race dyads may start at lower trust levels and thus be especially suited to reaping maximum rewards from the PD game; Tropp, 2008; Zolin & Gibbons, 2014), finding equal effectiveness of stakes for same- and cross-race dyads is noteworthy. Relatedly, the lack of observed moderation based on sharing a racial identity (or not) with one's partner suggests that group membership is unlikely to underly the observed effects, partially addressing the possibility that shared student identity (as undergraduates at the same university) could be impacting the results or limit generalizability. Despite many societal forces that could lead people to be distrusting of others outside their group (Foddy et al., 2009), our intervention's effects appear just as strong whether building trust within-group or bridging group boundaries.

Information or Payment: Which Factor Drives These Effects?

A key limitation of the present research is the joint manipulation of information and payment between the high- and low-stakes PD games. Confounding these two aspects of the PD games prevents definitively determining which aspect (information about partner choice or ability to win real money) drives the observed stakes effects. It is also important to note that because of this design, the moderation of stakes on partner cooperation's effects on trust is partially confounded due to the access to partner cooperation only being possible in the high-stakes (risky) PD game. Despite this limitation, some results and relevant literature point towards manipulation of information as the more influential factor. First, our initial use of different payoff matrices (which directly varied whether participants received \$7 or \$9 for cooperating in the high-stakes condition, as well as the corresponding temptation to defect for an \$3 or \$1) did not moderate effects of PD game stakes on cooperation levels or post-game outcomes, suggesting that participants were less concerned with the amount of money at stake than the

possibility of *visibly* taking advantage of their partner (See Appendix D). Moreover, for the high-stakes risky (vs. low-stakes safe) PD game, experiencers predicted beforehand that a typical undergraduate (though not their specific partner) would be *less* likely to defect (i.e., more likely to cooperate) in the first round and then reported afterward personally feeling *less* tempted to defect during the PD game. If payment were the key factor at play, one would expect defection to be more likely in the paid condition, not less. These findings also align with prior research finding that payment was relatively unimportant compared with communication and other structural factors for influencing cooperation (Jin et al., 2025).

Conclusions

Increasing trust between diverse coworkers, classmates, teammates, or neighbours—enabling them to work, play, and live together, even when their interests’ conflict—is a crucial goal for scientific inquiry and society (Putnam, 2007). The present research provides strong support that, rather than merely creating cooperative contexts to build trust between individuals, inducing risky interdependence—in our case a high-stakes PD game—in which cooperation (or lack thereof) has real social stakes is a robust and powerful method for building trust among strangers, even across racial group boundaries.

References

- Aiken, L. S., & West, S. G. (1991). Multiple regression: Testing and interpreting interactions. *Multiple Regression: Testing and Interpreting Interactions, Journal Article*.
- Allport, F. H. (1954). The structuring of events: Outline of a general theory with applications to psychology. *Psychological Review*, *61*(5), 281–303. <https://doi.org/10.1037/h0062678>
- Aron, A., & Aron, E. N. (1986). *Love and the expansion of self: Understanding attraction and satisfaction* (pp. x, 172). Hemisphere Publishing Corp/Harper & Row Publishers.
- Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, *63*(4), 596–612.
- Aron, A., Eberhardt, J. L., Davies, K., Bergsieker, H. B., & Wright, S. C. (unpublished manuscript). *Initial Test of a Social-Psychological Intervention to Improve Community Relations with Police*.
- Aron, A., Melinat, E., Aron, E. N., Vallone, R., & Bator, R. (1997). The experimental generation of interpersonal closeness: A procedure and some preliminary findings. *Personality and Social Psychology Bulletin*, *23*(4), 363–377. <https://doi.org/10.1177/0146167297234003>
- Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.
- Babbitt, L. G., & Sommers, S. R. (2011). Framing Matters: Contextual Influences on Interracial Interaction Outcomes. *Personality and Social Psychology Bulletin*, *37*(9), 1233–1244. <https://doi.org/10.1177/0146167211410070>
- Balliet, D., & Van Lange, P. A. M. (2013). Trust, conflict, and cooperation: A meta-analysis. *Psychological Bulletin*, *139*(Journal Article), 1090–1112.

- Berg, J., Dickhaut, J. W., & McCabe, K. A. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142.
- Bergsieker, H. B. (2012). Building, betraying, and buffering trust in interracial and same-race friendships. In *ProQuest Dissertations and Theses: Vol. Ph.D.* Princeton University.
- Bottom, W. P., Gibson, K., Daniels, S. E., & Murnighan, J. K. (2002). When talk is not cheap: Substantive penance and expressions of intent in rebuilding cooperation. *Organizational Science*, *13*(5), 497–513.
- Bourdieu, P. (1983). Forms of capital. In J. Richardon (Ed.), *Handbook of theory and research for the sociology of education* (1–Book, Section, pp. 214–258). Greenwood.
- Brewer, M. B. (2001). Ingroup identification and intergroup contact: When does ingroup love become outgroup hate? In R. Ashmore, L. Jussim, & D. Wilder (Eds.), *Social identity, intergroup conflict, and conflict reduction* (1–Book, Section, pp. 2–41). Oxford University Press.
- Chua, R. Y. J. (2013). The Costs of Ambient Cultural Disharmony: Indirect Intercultural Conflicts in Social Environment Undermine Creativity. *Academy of Management Journal*, *56*(6), 1545–1577. <https://doi.org/10.5465/amj.2011.0971>
- Chua, R. Y. J., Ingram, P., & Morris, M. W. (2008). From the Head and the Heart: Locating Cognition- and Affect-Based Trust in Managers' Professional Networks. *Academy of Management Journal*, *51*(3), 436–452. <https://doi.org/10.5465/amj.2008.32625956>
- Collins, M. G., Juvina, I., & Gluck, K. A. (2016). Cognitive Model of Trust Dynamics Predicts Human Behavior within and between Two Games of Strategic Interaction with Computerized Confederate Agents. *Frontiers in Psychology*, *7*, 49. <https://doi.org/10.3389/fpsyg.2016.00049>

- Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods, 14*(2), 165–176. <https://doi.org/10.1037/a0015565>
- Costafreda, S. (2009). Pooling fMRI data: Meta-analysis, mega-analysis and multi-center studies. *Frontiers in Neuroinformatics, 3*. <https://doi.org/10.3389/neuro.11.033.2009>
- Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*(4), 349–354. <https://doi.org/10.1037/h0047358>
- Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods, 14*(2), 81–100. <https://doi.org/10.1037/a0015914>
- Dovidio, J., Gaertner, S., Kawakami, K., & Hodson, G. (2002). Why can't we just get along? Interpersonal biases and interracial distrust. *Cultural Diversity and Ethnic Minority Psychology, 8*(Journal Article), 88–102.
- Drigotas, S. M., Rusbult, C. E., & Verette, J. (1999). Level of commitment, mutuality of commitment, and couple well-being. *Personal Relationships, 6*(3), 389–409. <https://doi.org/10.1111/j.1475-6811.1999.tb00199.x>
- Dunning, D., Anderson, J. E., Schlösser, T., Ehlebracht, D., & Fetchenhauer, D. (2014). *Trust at zero acquaintance: A matter of respect, not expectation*.
- Effron, D. A., & Miller, D. T. (2011). Reducing exposure to trust-related risks to avoid self-blame. *Personality and Social Psychology Bulletin, 37*(2), 181–192. <https://doi.org/10.1177/0146167210393532>

- Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global Evidence on Economic Preferences*. *The Quarterly Journal of Economics*, *133*(4), 1645–1692. <https://doi.org/10.1093/qje/qjy013>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Fetchenhauer, D., & Dunning, D. (2009). Do people trust too much or too little? *Journal of Economic Psychology*, *30*(3), 263–276.
- Fetchenhauer, D., & Dunning, D. (2012). Betrayal aversion versus principled trustfulness—How to explain risk avoidance and risky choices in trust games. *Journal of Economic Behavior & Organization*, *81*(2), 534–541.
- Fiedler, K., Harris, C., & Schott, M. (2018). Unwarranted inferences from statistical mediation tests – An analysis of articles published in 2015. *Journal of Experimental Social Psychology*, *75*, 95–102. <https://doi.org/10.1016/j.jesp.2017.11.008>
- Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178. <https://doi.org/10.1007/s10683-006-9159-4>
- Fischer, D. G., & Fick, C. (1993). Measuring Social Desirability: Short Forms of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement*, *53*(2), 417–424. <https://doi.org/10.1177/0013164493053002011>
- Foddy, M., Platow, M., & Yamagishi, T. (2009). Group-based trust in strangers: The role of stereotypes and expectations. *Psychological Science*, *20*(Journal Article), 419–422.

- Government of Canada, S. C. (2015, May 20). *The Daily — Study: Trends in social capital in Canada, 2003, 2008 and 2013*. <https://www150.statcan.gc.ca/n1/daily-quotidien/150520/dq150520d-eng.htm>
- Government of Canada, S. C. (2023, April 19). *Trust in neighbours*. <https://www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2023022-eng.htm>
- Government of Canada, S. C. (2024, May 16). *General trust in others by gender and other selected sociodemographic characteristics*. <https://www150.statcan.gc.ca/t1/tb11/en/tv.action?pid=4510009901>
- Hardin, R. (2002). *Trust and trustworthiness*. Russell Sage Foundation.
- Hart, K. (1988). Trust: Making and breaking cooperative relations. In D. Gambetta (Ed.), *Kinship, contract as trust: The economic organisation of migrants in an African city slum* (1–Book, Section, pp. 176–193). Basil Blackwell.
- Hewstone, M., Kenworthy, J. B., Cairns, E., Tausch, N., Hughes, J., Tam, T., & et al. (2008). Stepping stones to reconciliation in Northern Ireland: Intergroup contact, forgiveness, and trust. In A. Nadler, T. E. Malloy, & J. D. Fischer (Eds.), *The social psychology of intergroup reconciliation* (1–Book, Section, pp. 199–226). Oxford University Press.
- Holland, S. J., Shore, D. B., & Cortina, J. M. (2017). Review and recommendations for integrating mediation and moderation. *Organizational Research Methods*, 20(4), 686–720. <https://doi.org/10.1177/1094428116658958>
- Holmes, J. G. (1981). The exchange process in close relationships. In M. J. Lerner & S. C. Lerner (Eds.), *The Justice Motive in Social Behavior* (pp. 261–284). Springer US. https://doi.org/10.1007/978-1-4899-0429-4_12

- Holmes, J. G. (2002). Interpersonal expectations as the building blocks of social cognition: An interdependence theory perspective. *Personal Relationships*, 9(Journal Article), 1–26.
- Insko, C. A., & Schopler, J. (1998). Differential distrust of groups and individuals. In C. Sedikides, J. Schopler, & C. A. Insko (Eds.), *Intergroup cognition and intergroup behavior: Applied social research* (1–Book, Section, pp. 75–107). Erlbaum.
- Jin, S., Spadaro, G., & Balliet, D. (2025). Institutions and cooperation: A meta-analysis of structural features in social dilemmas. *Journal of Personality and Social Psychology*, 129(2), 286–312. <https://doi.org/10.1037/pspi0000474>
- Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, 68(1), 601–625. <https://doi.org/10.1146/annurev-psych-122414-033702>
- Kelley, H. H., & Thibaut, J. W. (1978). *Interpersonal relations*. Wiley.
- Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. Guilford Press.
- Kopelman, S., & Berkel, G. (2012). Kukui nuts. *Evanston, IL: Dispute Resolution Center*. <https://new.negotiationexercises.com/wp-content/uploads/2023/01/Kukui-Nuts-2020-Webinar.pdf>
- Korzinski, D. (2022, November 7). Democracy in North America: Significant segments in Canada, U.S. open to authoritarian leadership. *Angus Reid Institute*. <https://angusreid.org/democracy-and-authoritarianism-canada-usa/>
- Lewicki, R. J., Tomlinson, E. C., & Gillespie, N. (2006). Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. *Journal of Management*, 32(6), 991–1022.

- Lount, R. B. J., Zhong, C.-B., Sivanathan, N., & Murnighan, J. K. (2008). Getting off on the wrong foot: The timing of a breach and the restoration of trust. *Personality and Social Psychology Bulletin*, 34(12), 1601–1612.
- Marsden, P. V. (2011). Survey methods for network data. *The SAGE Handbook of Social Network Analysis*, 25, 370–388.
- Matzat, U., & Snijders, C. (2010). Does the online collection of ego-centered network data reduce data quality? An experimental comparison. *Social Networks*, 32(2), 105–111.
<https://doi.org/10.1016/j.socnet.2009.08.002>
- Montoya, A. K. (2024). Combining statistical and causal mediation analysis. In H. T. Reis, T. West, & C. M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (3rd ed., pp. 622–652). Cambridge University Press.
<https://doi.org/10.1017/9781009170123.026>
- Murnighan, J. K. (1991). *The dynamics of bargaining games*. Prentice Hall.
- Murray, S. L., & Holmes, J. G. (2009). The architecture of interdependent minds: A motivation-management theory of mutual responsiveness. *Psychological Review*, 116(4), 908–928.
<https://doi.org/10.1037/a0017015>
- Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, 45(4), 867–872. <https://doi.org/10.1016/j.jesp.2009.03.009>
- Phan, M. B. (2008). We're all in this together: Context, contacts, and social trust in Canada. *Analyses of Social Issues and Public Policy*, 8(Journal Article), 23–51.
- Putnam, R. D. (1993). *Making democracy work*. Princeton University Press.

- Putnam, R. D. (2000). Bowling alone: The collapse and revival of American community. *The Social Science Journal*, 39(Journal Article), 541.
- Putnam, R. D. (2007). E pluribus unum: Diversity and community in the twenty-first century The 2006 Johan Skytte Prize Lecture. *Scandinavian Political Studies*, 30(2), 137–174.
<https://doi.org/10.1111/j.1467-9477.2007.00176.x>
- Rapoport, A. (1967). A note on the index of cooperation for prisoner's dilemma. *Journal of Conflict Resolution*, 11(Journal Article), 100–103.
- Reis, H. T., & Shaver, P. (1988). Intimacy as an interpersonal process. In H. T. Reis & P. Shaver (Eds.), *Handbook of personal relationships: Theory, research, and interventions* (pp. 367–389). John Wiley & Sons.
- Rijnhart, J. J. M., Lamp, S. J., Valente, M. J., MacKinnon, D. P., Twisk, J. W. R., & Heymans, M. W. (2021). Mediation analysis methods used in observational research: A scoping review and recommendations. *BMC Medical Research Methodology*, 21(1), 226.
<https://doi.org/10.1186/s12874-021-01426-3>
- Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In *Advances in experimental social psychology* (Vol. 10, pp. 173–220). Elsevier.
<https://www.sciencedirect.com/science/article/pii/S0065260108603573>
- Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management. The Academy of Management Review*, 23(3), 393–404.
- Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science (Washington)*, 277(5328), 918–928.

- Sears, D. O. (1983). The person-positivity bias. *Journal of Personality and Social Psychology*, 44(2), 233–250.
- Sherif, M. (1966). *In common predicament: Social psychology of intergroup conflict and cooperation*. Houghton Mifflin.
- Simpson, J. A. (2007). Psychological foundations of trust. *Current Directions in Psychological Science*, 16(5), 264–268. <https://doi.org/10.1111/j.1467-8721.2007.00517.x>
- Smeesters, D., Warlop, L., Van Avermaet, E., Corneille, O., & Yzerbyt, V. (2003). Do not prime hawks with doves: The interplay of construct activation and consistency of social value orientation on cooperative behavior. *Journal of Personality and Social Psychology*, 84(5), 972–987. <https://doi.org/10.1037/0022-3514.84.5.972>
- Sparks, A., Burleigh, T., & Barclay, P. (2016). We can see inside: Accurate prediction of Prisoner's Dilemma decisions in announced games following a face-to-face interaction. *Evolution and Human Behavior*, 37(3), 210–216. <https://doi.org/10.1016/j.evolhumbehav.2015.11.003>
- Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, 89(6), 845–851. <https://doi.org/10.1037/0022-3514.89.6.845>
- Sung, Y. J., Schwander, K., Arnett, D. K., Kardia, S. L. R., Rankinen, T., Bouchard, C., Boerwinkle, E., Hunt, S. C., & Rao, D. C. (2014). An Empirical Comparison of Meta-analysis and Mega-analysis of Individual Participant Data for Identifying Gene-Environment Interactions. *Genetic Epidemiology*, 38(4), 369–378. <https://doi.org/10.1002/gepi.21800>

- Toosi, N. R., Babbitt, L. G., Ambady, N., & Sommers, S. R. (2012). Dyadic interracial interactions: A meta-analysis. *Psychological Bulletin*, *138*(1), 1–27.
<https://doi.org/10.1037/a0025767>
- Tropp, L. R. (2008). The role of trust in intergroup contact: Its significance and implications for improving relations between groups. In U. Wagner, L. R. Tropp, G. Finchilescu, & C. Tredoux (Eds.), *Improving intergroup relations: Building on the legacy of Thomas F. Pettigrew* (1–Book, Section, pp. 91–106). Blackwell Publishing.
- Van Lange, P. A. M., De Bruin, E. M. N., Otten, W., & Joireman, J. A. (1997). Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. *Journal of Personality and Social Psychology*, *73*(4), 733–746.
<https://doi.org/10.1037/0022-3514.73.4.733>
- Van Lange, P. A. M., & Rusbult, C. E. (2012). Interdependence theory. In P. Van Lange, A. Kruglanski, & E. Higgins (Eds.), *Handbook of Theories of Social Psychology* (1–Book, Section, pp. 251–273). Sage.
- von der Lippe, H., & Gamper, M. (2017). Drawing or tabulating ego-centered networks? A mixed-methods comparison of questionnaire vs. visualization-based data collection. *International Journal of Social Research Methodology*, *20*(5), 425–441.
<https://doi.org/10.1080/13645579.2016.1227649>
- Vorauer, J. D. (2006). An information search model of evaluative concerns in intergroup interaction. *Psychological Review*, *113*(4), 862–886. <https://doi.org/10.1037/0033-295X.113.4.862>
- Wellman, B. (1979). The Community Question: The Intimate Networks of East Yorkers. *American Journal of Sociology*, *84*(5), 1201–1231. <https://doi.org/10.1086/226906>

- Wright, S. C., Brody, S. A., & Aron, A. (2005). Intergroup contact: Still our best hope for improving intergroup relations. In C. S. Crandall & M. Schaller (Eds.), *Social psychology of prejudice: Historical perspectives* (1–Book, Section). Lewinian Press.
- Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, *51*(1), 110–116. <https://doi.org/10.1037/0022-3514.51.1.110>
- Zhao, M., Li, Y., Lin, J., Fang, Y., Yang, Y., Li, B., & Dong, Y. (2024). The Relationship Between Trust and Well-Being: A Meta-Analysis. *Journal of Happiness Studies*, *25*(5), 1–26. <https://doi.org/10.1007/s10902-024-00737-8>
- Zolin, R., & Gibbons, D. E. (2014). How Emergent Roles and Structures Create Trust in Hastily Formed Interorganizational Teams. *SAGE Open*, *4*(2), 2158244014533555. <https://doi.org/10.1177/2158244014533555>

Appendix A: Pilot Study Results

Pilot study. Prior to Study 1, a small ($N = 28^7$) pilot experiment open to all undergraduates was conducted to test and refine the procedure, plus ascertain whether the six most common racial groups on campus were perceived to have intergroup trust gaps. The 28 participants who self-identified as White (14), East Asian (7), South Asian (5), Latino (1), or Middle Eastern (1) were retained for analysis. After experiencing and giving feedback on the Study 1a tasks, pilot participants answered questions assessing which groups their racial ingroup trusted and which groups trusted their ingroup.⁸ Specifically, participants estimated the extent to which most undergraduates at their university “think that [participant’s race] students trust [White/Black/Latino/East Asian/South Asian/Middle Eastern] students here” and “think that [White/Black/Latino/East Asian/South Asian/Middle Eastern] students trust [participant’s race] students here” on a scale “from 0 (not at all) to 100 (completely).” The six outbound trust items (whom does your group trust?) preceded the six inbound items (who trusts your group?). Intragroup trust was computed by averaging the two items that included a given participant’s own race as both the trusting and trusted group. A general intergroup trust measure was computed by averaging all items about the ingroup trusting and being trusted by the other five racial groups. The final specific measure of intergroup trust was computed by averaging all items about trust between Whites and East Asians (the only groups with $n > 5$).

A repeated-measures ANOVA revealed higher estimates for how much students trusted—and were trusted by—racial ingroup members ($M = 85.38$, $SE = 3.02$) versus all

⁷ Excludes 2 participants dropped due to a technical error and 1 who declined to identify with a racial group.

⁸ Parallel questions about personal likelihood estimates that participants from various racial groups would choose “X” in the PDG when paired with ingroup or outgroup partners failed to yield meaningful results due to ceiling effects.

outgroup members ($M = 75.27, SE = 3.34$), $F(1, 27) = 8.81, p = .006, \eta^2_p = .25, d = 0.60$.⁹

Including participant race in a mixed-factorial ANOVA initially proved problematic due to low n and significant heterogeneity of variance across groups, Levene's $F(2, 23) = 8.44, p = .002$.

Variance could not be computed for groups with a single participant, and inspection revealed the highest SDs among South Asians for both trust measures; descriptively, only South Asians reported less intragroup than intergroup trust. A simplified model retaining Whites and East Asians confirmed higher levels of intragroup ($M = 84.41, SE = 2.40$) than specific intergroup trust ($M = 68.71, SE = 3.96$), $F(1, 19) = 25.19, p < .001, \eta^2_p = .57, d = 1.05$. Participant race did not significantly qualify this effect, $F(1, 19) = 1.89, p = .185, \eta^2_p = .09$, but Whites estimated higher overall trust levels ($M = 84.41, SE = 2.40$) than did East Asians ($M = 68.71, SE = 3.96$), $F(1, 19) = 6.65, p = .018, \eta^2_p = .26, d = 1.19$. Even in this small sample, which limits the precision of group-specific estimates, a striking intergroup trust gap emerged between White and East Asian students, the two largest groups on this campus.

⁹ Unless otherwise noted, for within-participant effects Cohen's d_{av} (based on condition SDs) is reported, not d_z (based on SDs of condition differences), to facilitate interpretation and comparison with between-participant effects.

Appendix B: Additional Materials and Measures

Friendship Network (Studies 2 & 3)

Participants' social networks were assessed using an ego-centric name generator (Marsden, 2011; Wellman, 1979) and a "Friendship Network" sociomatrix commonly used in management/network science (e.g., Chua, 2013; Chua et al., 2008; Matzat & Snijders, 2010; von der Lippe & Gamper, 2017). Participants first nominated and reported ties between friends, then reported how compatible each friend would be with their partner and how likely they would introduce each friend to their partner. Finally, participants reported the gender, race, and academic major/area of each friend.

Participants were first asked to "list the first names of your 10 closest friends who attend [this university], with whom you spend the most time," excluding romantic partners. Then, using a square adjacency matrix with friends' names listed as row and column labels, participants indicated whether each person on the row label saw each person on the column label as a friend (up to 90 potential friendship ties, excluding reflexive ties on the diagonal).

Friendship Network Introductions

Participants rated the extent to which each friend listed in the friendship network would be compatible with their partner on a scale from 1 (*Not at all*) to 5 (*Extremely*). Participants rated the likelihood that they would introduce each friend in their network to their partner on a scale from 1 (*Definitely no*) to 4 (*Definitely yes*).

Generalizing PD Game to Others

Participants in Study 3 were given another exploratory measure after completing the friendship network task. They were given a profile to make predictions about behaviour in a PD game. They completed this task five times, in which we varied their partner's gender and race.

Discussion Topic Task

Participants completed an exploratory task in which they ranked current societal or student issues (i.e., racism, diversity, environment, tuition, degrees) they would like to discuss with their partner. This task took place immediately prior to debriefing for Study 1, 4, 5, but in Studies 2 and 3 participants completed this task immediately after the interpersonal assessment scale (before the friendship network task). This exploratory measure failed to produce clear results (most students avoided intergroup topics), so it is not included in our analyses.

Negotiation Task Details

Participants played the roles of representatives of two pharmaceutical companies competing to purchase kukui nuts, a scarce and desirable resource, from a third-party seller. Participants were told that both companies want to obtain as many nuts as possible, and was required to negotiate over how the nuts would be split between the companies. The case had integrative potential in that both companies needed different parts of the nuts: one company only wanted the shell, whereas the other only wanted the seed. Thus, disclosure of critical information (e.g., why they wanted the nuts) could allow both parties to share the nuts by only taking the parts they need, resulting in both companies obtaining the maximum quantity of available nuts.

Participants were given five minutes to read over individual role sheets explaining the case, the background of their companies, how their companies were planning to use the nuts, and their budget for purchasing the nuts. Participants were told that they may discuss any information, but to refrain from showing their role sheets to their partner. Participants were then given ten minutes to negotiate.

Appendix C: Scale Items and Reliabilities

Table C1: Reliabilities and Individual Items for Interpersonal Assessment

	Reliability	
	E	F
Trust, 18-item scale		
1 (strongly disagree); 7 (strongly agree)		
Sometimes I worry that [Partner's name] may take advantage of me. (R)	$\alpha: .89$	$\alpha: .90$
I can confide in [Partner's name] and know that he/she would not discuss my concerns with others.		
I can count on [Partner's name] to be concerned about my well-being.		
Even if it required a personal sacrifice, [Partner's name] would support me when I needed help.		
If I asked [Partner's name] to call me at a certain time, I could count on receiving the call.		
In most matters, I trust [Partner's name] completely.		
[Partner's name] is usually dependable, especially for things that really matter to me		
Cultural trust, 4-item scale ^a	$\alpha: .66$	$\alpha: .73$
1 (strongly disagree); 7 (strongly agree)		
[Partner's name] would not second-guess my reaction if I found someone's comments culturally offensive.		
If I said something that might seem culturally insensitive, [Partner's name] would give me the benefit of the doubt.		
[Partner's name] would be supportive if I avoided an activity because of my upbringing or background.		
If we disagreed over a cultural difference, [Partner's name] would respect my opinions and position.		
Liking, 6-item scale	$\alpha: .83$	$\alpha: .79$
1 (strongly disagree); 7 (strongly agree)		
Generally speaking, I really like [Partner's name].		
I sometimes dislike [Partner's name]. (R)		
I sometimes find [Partner's name] irritating or unpleasant. (R)		
My interactions with [Partner's name] are typically pleasant and enjoyable.		
I enjoy being around [Partner's name].		
I sometimes try to avoid spending time with [Partner's name]. (R)		

Note. E = Experiencers. F = Forecasters. ^a dropped fifth item, ^b Studies 1, 4 and 5 participants SDS was obtained through pre-screen. Items marked with (R) are reverse-scored.

Table C2: Reliabilities and Individual Items for Covariates

Covariates	Reliability	
	E	F
General trustfulness (5 items)		
1 (Strongly disagree); 4 (Neutral); 7 (Strongly agree)	$\alpha = .62$	$\alpha = .64$
Most people are basically honest.		
People are always interested only in their own welfare		
Most people will respond in kind when they are trusted by others		
Most people are basically good and kind.		
People usually do not trust others as much as they say they do.		
Social Desirability Scale ^a (6 items)	$\alpha = .59$	$\alpha = .54$
True; False		
I have never intensely disliked anyone.		
I sometimes feel resentful when I don't get my way. (R)		
There have been times when I felt like rebelling against people in authority even though I knew they were right. (R)		
I am always courteous, even to people who are disagreeable.		
There have been times when I was quite jealous of the good fortune of others.(R)		
I am sometimes irritated by people who ask favors of me. (R)		
<i>Note.</i> E = Experiencers. F = Forecasters. ^a Studies 1, 4 and 5 participants SDS was obtained through pre-screen. Items marked with (R) are reverse-scored.		

Table C3: Pre- and Post-game Slider Items and Reliabilities (Experiencers)

Pre- and post-game sliders	Reliability	
	Pre-game	Post-game
Not at all; extremely		
Trust (2 items)	$\alpha = .83$	$\alpha = .93$
How much do you TRUST [Partner's name]?		
How much do you think you can RELY on [Partner's name]?		
Liking (2 items)	$\alpha = .85$	$\alpha = .91$
How much do you LIKE [Partner's name]?		
How much do you ENJOY being around [Partner's name]?		
Closeness (2 items)	$\alpha = .76$	$\alpha = .84$
How CLOSE do you feel to [Partner's name]?		
How WELL do you know [Partner's name]?		
Gratitude and relief (2 items)	—	$\alpha = .70$
1 (not at all); 7 (very much) (Study 2 & 3)		
Grateful		
Relieved		

Note. No numbers were displayed to participants for the sliders.

Table C4: Reliabilities and Individual Items for Negotiation Variables (Studies 2 & 3)

Negotiation variables	Reliability
1 (strongly disagree); 7 (strongly agree)	
Trust (3 items)	α : .77
I trusted [Partner's name] during the negotiation	
During the negotiation I worried that [Partner's name] might take advantage of me. (R)	
During the negotiation I felt that [Partner's name] tried to manipulate me (R)	
Info sharing (2 items)	α : .79
I shared information fully and openly with [Partner's name].	
[Partner's name] shared information fully and openly with me.	
Satisfaction with outcomes (2 items)	α : .76
How satisfied are you with your own outcome—i.e., the extent to which the terms of your agreement (or lack of agreement) benefit you?	
How satisfied are you with the balance between your own outcome and [Partner's name]'s outcome(s)?	
Partner treatment during negotiation (2 items)	α : .87
Did [Partner's name] listen to your concerns?	
Did [Partner's name] try to accommodate your wishes, opinions, or needs?	
Comfort during negotiation (2 items)	α : .49
The conversation ran smoothly without any uncomfortable silences.	
I was nervous talking to [Partner's name]. (R)	
Understanding during negotiation (2 items)	α : .84
Throughout the negotiation, I understood [Partner's name] well.	
Throughout the negotiation, [Partner's name] understood me well.	
Partner warmth and competence during negotiation (2 items)	α : .79
[Partner's name] was very warm, friendly and likeable.	
[Partner's name] was very competent, intelligent, and skilled.	

Note. Items marked with (R) are reverse-scored.

Appendix D: Analysis of Moderation by PD Game Matrix

Despite expectations that the 7/3 (vs. the 9/1) matrix would be especially conducive to building trust, we see no evidence of the matrix configuration moderating any stakes effects on outcomes. When predicting experiencer's confidence that a typical undergraduate would cooperate in the first round we see the only significant stakes-by-matrix interaction, such that in the 9/1 matrix experiencers are more confident in the risky (vs. safe) condition.

Table D1: Moderation of Stakes Effects by PD Game Matrix

Variable	Experiencers				Forecasters		
	Stakes effect		Stakes × matrix	Stakes × matrix	Stakes × Role × matrix	Stakes × matrix	Role × matrix
	9/1 matrix	7/3 matrix					
	<i>t</i>	<i>d</i>	<i>t</i>	<i>d</i>	<i>t</i>	<i>t</i>	<i>t</i>
Cooperation							
Confidence in partner R1 co-op	0.19	0.06	0.67	-0.19	-0.57	0.81	-0.68
Confidence in undergrad R1 co-op ^a	2.40*	0.86	0.61	-0.17	-2.24*	-1.07	-0.69
Probability partner cooperates 100%	0.64	-0.23	0.42	-0.12	0.23	-0.29	0.28
Own cooperation rate	2.55*	0.87	2.56*	0.71	-0.36	-0.59	0.11
Partner cooperation rate	2.49*	0.86	2.39*	0.67	-0.42	-0.40	0.01
PD game reactions							
Cared about partners choices	4.25***	1.44	4.81***	1.39	-0.26	-0.86	0.47
Tempted to defect	1.45	-0.50	0.18	-0.05	1.01	0.77	0.58
Worried partner would defect	0.41	-0.14	1.14	0.32	1.03	0.23	0.96
Trust							
Trust (single-item)	4.83***	1.63	4.82***	1.38	-0.71	-0.05	-0.39
Reliance (single-item)	3.52***	1.22	4.20***	1.18	-0.08	0.48	-0.31
Trust (PD-game impact)	4.79***	1.63	7.11***	2.00	0.77	-0.57	0.98
Trust (perceived change)	5.53***	1.85	5.72***	1.63	-0.68	0.06	-0.56
Trust (post-game slider)	4.17***	1.45	5.47***	1.58	0.21	—	—
Liking							
Liking (single-item)	1.87 [†]	0.62	3.00**	0.84	0.44	0.41	0.09
More pleasant (perceived change)	5.60***	1.94	6.35***	1.78	-0.34	-0.02	-0.32
Liking (post-game slider)	0.85	0.29	1.08	0.31	0.02	—	—
Relationship							
Closeness (PD-game impact)	3.35**	1.16	5.03***	1.41	0.58	0.83	-0.03
Closer (perceived change)	5.28***	1.84	3.48***	0.98	-1.90 [†]	0.55	-1.80 [†]
Closeness (post-game slider)	2.40*	0.85	3.75***	1.07	0.51	—	—
IOS (post-game)	3.03**	1.04	1.50	0.42	-1.41	—	—
Stronger (perceived change)	4.59***	1.62	4.64***	1.30	-0.63	0.49	-0.84
Better (perceived change)	4.47***	1.52	5.61***	1.60	0.07	-0.02	-0.11

Note. Boldface indicates significant effects. R1 = Round One. IOS = inclusion of other in self. ^a Confidence in typical undergraduate Round One cooperation was not measured in Study 5.

Slider measures and IOS were only collected for experiencers. *** $p < .001$. ** $p < .01$. * $p < .05$. [†] $p < .1$.

Appendix E: Analysis of Secondary Variables

Looking at the interpersonal assessment scale, experiencers in the risky (vs. safe) condition showed meaningful gains on trust ($b = 0.31, p < .001, d = 0.56$), and liking ($b = 0.23, p = .002, d = 0.38$). Forecasters completing the risky (vs. safe) PD game had the same pattern of results, greater trust ($b = 0.18, p = .044, d = 0.23$), and liking ($b = 0.20, p = .040, d = 0.23$). There were no significant differences for the cultural trust scale (all $t_s < 1.40$). Across all three subscales there were no significant role-by-stakes interactions (all $t_s < 1.21$). See Table E1 for descriptives and nesting for experiencers.

Table E1: Interpersonal Assessment Measures by Stakes and Role

Outcome	ICC	Experiencers				Forecasters				Role × Stakes <i>t</i>
		Safe Mean (SD)	Risky Mean (SD)	Stakes		Safe Mean (SD)	Risky Mean (SD)	Stakes		
				<i>t</i>	<i>d</i>			<i>t</i>	<i>d</i>	
Cooperation										
Trust (18-item)	.12	4.76 (0.75)	5.07 (0.72)	4.80***	0.56	4.19 (0.78)	4.36 (0.80)	2.02*	0.23	1.20
Liking (6-item)	.27	5.53 (0.81)	5.76 (0.73)	3.20**	0.38	4.31 (0.81)	4.50 (0.92)	2.06*	0.23	0.22
Cultural trust (4-item)	.06	4.58 (0.80)	4.68 (.80)	1.39	0.16	4.32 (0.92)	4.42 (0.91)	1.08	0.12	-0.27

Note. All $N_s = 892$, as interpersonal assessments not measured for study 4. ICC shows the nesting structure of the data for experiencers within dyads.

Appendix F: Behavioural Coding of Study 3 Experiencers

In Study 3 the negotiation following their (risky or safe) PD game was videorecorded, with a separate camera trained on each participant. Research assistants separately coded the verbal and non-verbal components of participants' behaviour. Participants displayed less awkwardness in the risky (vs. safe) condition, ($b = -0.24$, $p = .027$, $d = 0.40$), but otherwise no main effects of stakes emerged (see Table F1). Same-race (vs. cross-race) dyads showed marginally more nonverbal engagement ($b = 0.16$, $p = .074$, $d = 0.33$), and less nonverbal anxiousness ($b = -0.19$, $p = .025$, $d = 0.42$), on average. The interaction of stakes by dyad race was marginal for behaviourally coded smiling ($b = -0.39$, $t = -1.78$, $p = .078$), such that the effect of stakes on smiling was trending towards significance in the cross-race ($b = 0.25$, $p = .107$, $d = 0.41$), but was in the opposite direction (and not trending towards significance) in the same-race condition ($b = -0.14$, $p = .378$, $d = 0.24$).

Table F1: Stakes Main Effects on Negotiation Behaviour

Outcomes	Safe M (SD)	Risky M (SD)	Stakes Main effect
Non-verbal			
Engagement	4.81 (0.67)	4.96 (0.46)	0.07 (0.04)
Anxious	3.05 (0.66)	2.99 (0.59)	-0.02 (0.04)
Smiling	4.73 (0.78)	4.80 (0.60)	0.03 (0.06)
Nodding	4.12 (0.73)	4.12 (0.69)	0.00 (0.05)
Eye contact	4.64 (0.77)	4.78 (0.58)	0.06 (0.05)
Leaning toward partner	4.28 (0.99)	4.27 (0.99)	-0.01 (0.07)
Awkwardness	4.19 (0.62)	4.26 (0.61)	-0.12* (0.05)
Fidgeting	4.56 (0.54)	4.60 (0.53)	-0.01 (0.05)
Friendly	4.72 (0.64)	4.76 (0.50)	0.01 (0.05)
Cooperative	4.75 (0.53)	4.83 (0.38)	0.04 (0.04)
Dominant	3.80 (0.52)	3.87 (0.57)	0.03 (0.04)
Verbal			
Amount of talking	4.63 (0.76)	4.60 (0.77)	0.04 (0.03)
Disclosing info	3.11 (0.74)	2.86 (0.67)	0.06 (0.07)
Honesty	4.04 (0.89)	4.17 (0.76)	0.01 (0.04)
Engagement	5.00 (0.65)	5.08 (0.52)	0.03 (0.04)
Anxious	2.88 (0.64)	2.74 (0.56)	-0.07 (0.04)
Friendly	4.67 (0.61)	4.83 (0.52)	0.07 (0.05)
Cooperative	4.81 (0.57)	4.89 (0.51)	0.04 (0.04)
Dominant	3.96 (0.67)	3.99 (0.63)	0.01 (0.04)

Note. Boldface indicates significant effects. *N*s range from 226-229.

Appendix G: Moderation Analyses

We did not have any specific hypotheses about moderation by demographic variables, however to ensure we did not miss any potential reversals of the stakes effects based on social desirability, general trust, socio-economic status (SES), gender or participant race, we ran models with the stakes-by-role-moderator interaction term. For variables that were only measured in experiencers, we simply included the stakes-by-moderator term.

There were few significant effects following no particular pattern, and given our $\alpha = .05$, finding seven significant effects after running 135 tests is about exactly what you would expect if there is no underlying association between variables (see Table G1). However, because of our specific interest on trust, we break down the three-way interaction of stakes-by-role-by-SES, to ensure that our intervention is effective for people of varying levels of SES. Breaking down the three-way into the simple interactions within experiencers and forecasters, we see a significant interaction of SES and stakes for experiencers ($b = -0.20, t = -2.08, p = .038$), but not forecasters ($b = 0.08, t = 1.11, p = .267$). This simple interaction was such that participants lower (vs. higher) in SES experienced a larger gain in trust in the risky (vs. safe) PD game, however regardless of SES the effect of the risky (vs. safe) PD game was still positive and significant ($ps < .001$).

Descriptively, the other three-way interactions are as such: Higher general trust for experiencers (not forecasters) predicts greater positive effects of risky (vs. safe) on predicting that one's partner will cooperate in the first round, higher SES for forecasters (not experiencers) predicts greater positive effects of risky (vs. safe) on predicted partner cooperation rate, lower social desirability for experiencers (not forecasters) predicts greater positive effects of risky (vs. safe) on caring about one's partner's choices, being male predicted perceiving marginally greater positive effects of the risky (vs. safe) PD game on pleasantness for experiencers (not forecasters), both (experiencers and forecasters) simple interactions of gender on the stakes effects on perceiving a better relationship after the PD game were only trending ($ps > .16$), and higher SES experiencers reported greater positive effects of risky (vs. safe) on information sharing during the negotiation.

Table G1: Demographic Variables Moderation of Stakes Effects

Variable	SDS	Gentrust	SES	Gender	Race
All Participants					
	Stakes × Role × Moderator				
	<i>t</i>	<i>t</i>	<i>t</i>	<i>t</i>	<i>t</i>
Cooperation					
Confidence in partner R1 co-op	-0.97	2.34*	-1.03	-0.74	0.02
Confidence in undergrad R1 co-op ^a	-0.58	1.28	-0.36	1.04	-1.25
Probability partner cooperates 100%	-1.13	0.28	-0.71	-0.01	0.37
Own cooperation rate	-1.40	0.87	-1.24	-1.15	0.31
Partner cooperation rate	-0.63	-0.06	-2.29*	-1.86 [†]	-0.97
PD game reactions					
Cared about partners choices	-2.63**	0.23	0.05	0.50	0.24
Tempted to defect	0.25	0.93	0.19	0.51	0.57
Worried partner would defect	0.40	1.44	-0.42	0.04	1.71 [†]
Trust					
Trust (single-item)	-0.43	0.53	-3.01**	-0.20	1.20
Reliance (single-item)	-0.91	0.56	-1.18	1.55	1.77 [†]
Trust (PD-game impact)	0.44	1.44	-1.40	-1.29	0.26
Trust (perceived change)	-1.22	0.16	-1.78 [†]	0.97	0.40
Liking					
Liking (single-item)	-1.11	0.95	-1.22	0.82	0.65
More pleasant (perceived change)	-0.59	1.21	-0.26	2.09*	-0.08
Relationship					
Closeness (PD-game impact)	0.73	1.68 [†]	-0.70	0.15	0.24
Closer (perceived change)	-1.04	0.76	-0.03	0.33	0.03
Stronger (perceived change)	-0.35	0.22	-0.20	0.92	1.53
Better (perceived change)	0.26	1.47	-0.10	1.97*	1.30
Experiencer-only					
	Stakes × Moderator				
	<i>t</i>	<i>t</i>	<i>t</i>	<i>t</i>	<i>t</i>
Slider measures					
Trust (post-game slider)	-0.63	-0.18	-0.80	0.27	1.73 [†]
Liking (post-game slider)	1.02	-0.01	1.08	-0.21	-1.20
Closeness (post-game slider)	0.12	-0.81	0.49	1.32	-0.14
IOS (post-game)	0.01	0.79	0.16	-1.14	-0.49
Gratitude and relief	-0.53	0.84	-0.90	-0.63	0.20
Negotiation measures					
Info sharing	-0.09	0.99	2.30*	0.91	1.40
Trust during negotiation	-0.76	-0.36	-1.14	-0.11	0.46
Perceived change in trust	0.32	-0.44	0.09	-0.07	-0.09
More pleasant (perceived change)	0.51	0.54	-0.10	0.62	0.53

Note. Boldface indicates significant effects. R1 = Round One. IOS = inclusion of other in self. ^a Confidence in typical undergraduate Round One cooperation was not measured in Study 5. *** $p < .001$. ** $p < .01$. * $p < .05$. [†] $p < .1$.

We had expected some moderation of effects by social value orientation (SVO), however the pattern of (nonsignificant) effects extends here as well (see Table G2 for all tests of moderation, simple interactions, and simple effects for experiencers). There was only one significant three-way interaction for SVO, however both simple interactions were nonsignificant. The significant simple interaction on cooperation was such that participants with prosocial SVO reported a smaller effect of risky (vs. safe); however both prosocial and proself experiencers still experienced positive effects of the risky (vs safe) PD game. Proself forecasters predicted no stakes effect on caring about what their partner would choose during the PD game, whereas prosocial forecasters predicted that they would care more about their partner's choice in the risky (vs. safe) condition. Proself forecasters predicted a marginally positive impact of the risky (vs. safe) PD game on closeness, while prosocial forecasters did not predict any effect of the risky (vs. safe) PD game on closeness.

We had also expected moderation by dyad race, and yet the nonsignificant moderations continue in Table G3. The significant three-way interaction on confidence in partner Round One cooperation only yielded a marginal interaction for forecasters (vs. no significant interaction for experiencers), such that forecasters in same-race (vs. cross-race) dyads were marginally more confident that their partner would cooperate in the first round in the risky (vs. safe) PD game. However, experiencers in cross-race (vs. same-race) dyads predicted that a typical undergraduate would be more likely to cooperate in the risky (vs. safe) PD game. Forecasters in same-race (vs. cross-race) dyads additionally predicted that they would care more about their partners choices in the risky (vs. safe) PD game.

Table G2: Tests of Moderation by Social Value Orientation

Variable	Experiencers				Forecasters Stakes×		
	Stakes effect		Stakes × SVO	Stakes × SVO	Role × SVO		
	Proself	Prosocial					
	<i>t</i>	<i>d</i>	<i>t</i>	<i>d</i>	<i>t</i>	<i>t</i>	<i>t</i>
Cooperation							
Confidence in partner R1 co-op	0.67	0.11	0.03	0.00	-0.54	-0.29	-0.07
Confidence in undergrad R1 co-op ^a	1.12	0.18	1.93 [†]	0.27	0.24	-0.37	0.46
Probability partner cooperates 100%	0.90	0.15	0.89	-0.12	-1.26	0.85	-1.48
Own cooperation rate	6.58^{***}	1.78	4.88^{***}	0.63	-2.48*	0.11	-1.26
Partner cooperation rate	3.97^{***}	0.75	6.43^{***}	1.03	0.63	0.21	-0.02
PD game reactions							
Cared about partners choices	6.04^{***}	0.96	6.64^{***}	0.87	-0.99	-2.89^{**}	1.57
Tempted to defect	3.92^{***}	-0.69	3.26^{**}	-0.43	1.27	-0.04	0.87
Worried partner would defect	2.28*	-0.36	0.01	0.00	1.89 [†]	-0.99	1.95 [†]
Gratitude and relief	8.54^{***}	1.12	9.19^{***}	1.45	-1.49	—	—
Trust							
Trust (single-item)	8.71^{***}	1.48	9.47^{***}	1.24	-1.46	-1.49	0.39
Reliance (single-item)	7.22^{***}	1.20	7.18^{***}	0.95	-1.65 [†]	-1.87 [†]	0.47
Trust (PD-game impact)	8.62^{***}	1.48	11.44^{***}	1.48	-0.28	-0.80	0.45
Trust (perceived change)	10.48^{***}	1.66	12.66^{***}	1.67	-1.06	1.02	-1.52
Trust (post-game slider)	6.75^{***}	1.27	8.27^{***}	1.06	-0.63	—	—
Liking							
Liking (single-item)	5.09^{***}	0.86	5.00^{***}	0.66	-1.16	-0.67	-0.07
More pleasant (perceived change)	11.01^{***}	1.81	13.61^{***}	1.75	-0.89	0.39	-0.87
Liking (post-game slider)	3.75^{***}	0.62	4.65^{***}	0.62	-0.33	—	—
Relationship							
Closeness (PD-game impact)	7.89^{***}	1.38	8.16^{***}	1.05	-1.59	-2.21*	-0.67
Closer (perceived change)	9.01^{***}	1.42	9.37^{***}	1.20	-1.79 [†]	1.54	-2.34*
Closeness (post-game slider)	5.76^{***}	1.15	7.43^{***}	0.97	-0.28	—	—
IOS (post-game)	6.84^{***}	2.03	6.46^{***}	0.84	-1.81 [†]	—	—
Stronger (perceived change)	7.69^{***}	1.22	9.12^{***}	1.19	-0.82	0.52	-0.95
Better (perceived change)	8.42^{***}	1.33	10.70^{***}	1.41	-0.49	-0.18	-0.15
Negotiation							
Info sharing	2.10*	0.40	1.06	0.16	-1.20	—	—
Trust during negotiation	1.92 [†]	0.37	2.49*	0.39	-0.01	—	—
Perceived change in trust	0.23	0.04	1.18	0.18	0.57	—	—
More pleasant (perceived change)	0.25	0.05	1.42	0.22	0.67	—	—

Note. Boldface indicates significant effects. SVO = social value orientation. R1 = Round One. IOS = inclusion of other in self. ^a Confidence in typical undergraduate Round One cooperation was not measured in Study 5. Slider measures, IOS, and negotiation-related variables were only collected for experiencers. *** $p < .001$. ** $p < .01$. * $p < .05$. † $p < .1$.

Table G3: Tests of Moderation by Dyad Race

Variable	Experiencers				Forecasters		Stakes ×
	Stakes effect		Stakes ×		Stakes ×	Role ×	
	Cross-race	Same-race	Dyad Race	Dyad Race	Dyad Race	Dyad Race	
	<i>t</i>	<i>d</i>	<i>t</i>	<i>d</i>	<i>t</i>	<i>t</i>	<i>t</i>
Cooperation							
Confidence in partner R1 co-op	1.05	0.18	0.54	-0.09	-1.13	1.95 [†]	-2.34*
Confidence in undergrad R1 co-op ^a	2.79**	0.47	0.22	0.04	-1.84 [†]	—	—
Probability partner cooperates 100%	0.69	0.12	1.09	-0.18	1.26	-1.00	0.31
Own cooperation rate	5.15***	0.85	5.51***	0.92	0.19	-0.70	0.82
Partner cooperation rate	4.82***	0.79	5.70***	1.01	0.55	0.36	-0.04
PD game reactions							
Cared about partners choices	6.23***	1.03	6.05***	0.98	-0.20	2.45*	-2.22*
Tempted to defect	3.49***	-0.59	2.91**	-0.48	0.46	0.50	-0.22
Worried partner would defect	1.57	-0.26	0.07	-0.01	1.07	1.68 [†]	-0.89
Gratitude and relief	7.86***	1.59	9.12	1.81	0.82	—	—
Trust							
Trust (single-item)	9.02***	1.53	8.37***	1.39	-0.57	0.72	-0.94
Reliance (single-item)	7.56***	1.29	6.40***	1.05	-0.91	-0.12	-0.37
Trust (PD-game impact)	9.72***	1.61	10.21***	1.66	-0.22	1.34	-1.03
Trust (perceived change)	12.46***	2.21	10.58***	1.72	-1.47	-0.16	-0.61
Trust (post-game slider)	8.41***	1.42	6.51***	1.06	1.45	—	—
Liking							
Liking (single-item)	5.76***	0.99	3.55***	0.59	-1.63	-0.24	-0.69
More pleasant (perceived change)	12.80***	2.33	11.68***	1.90	-0.95	-0.30	-0.21
Liking (post-game slider)	4.54***	0.78	4.07***	0.66	-0.39	—	—
Relationship							
Closeness (PD-game impact)	7.76***	1.30	7.68***	1.23	-0.16	-0.30	0.05
Closer (perceived change)	9.38***	1.63	8.35***	1.36	-0.85	-0.49	-0.06
Closeness (post-game slider)	6.58***	1.11	6.56***	1.07	-0.11	—	—
IOS (post-game)	7.40***	1.24	5.77***	0.93	-1.26	—	—
Stronger (perceived change)	8.66***	1.49	7.83***	1.29	-0.70	-0.98	0.47
Better (perceived change)	10.40***	1.82	8.63***	1.43	-1.38	-1.45	0.48
Negotiation							
Info sharing	1.65	0.33	0.82	0.16	-0.60	—	—
Trust during negotiation	2.52*	0.50	1.66 [†]	0.33	-0.62	—	—
Perceived change in trust	1.05	0.21	0.56	0.11	-0.35	—	—
More pleasant (perceived change)	1.17	0.23	0.76	0.15	-0.29	—	—

Note. Boldface indicates significant effects. R1 = Round One. IOS = inclusion of other in self. ^a Confidence in typical undergraduate Round One cooperation was not measured in Study 5. Slider measures, IOS, and negotiation-related variables were only collected for experiencers. *** $p < .001$. ** $p < .01$. * $p < .05$. [†] $p < .1$.

Appendix H: Within-participant analysis of Study 4

Repeated measures ANOVAs with matrix (7/3 vs. 9/1) and stakes (risky vs. safe) were conducted for the Study 4 forecasters on our four trust operationalizations, single-item liking and perceived change in liking, and perceived change in and impact of the PD game on closeness. For all models, the interaction of stakes by matrix was not significant (all p s > .198). Stakes significantly predicted post-game reliance and PD game impact on trust, F s (1, 108) > 6.46, p s < .012, η_p^2 s > .06, but did not significantly predict post-game single-item trust or perceived change in trust (p s > .614). Stakes additionally predicted PD game impact on closeness, F (1, 108) = 11.33, p = .001, η_p^2 = .09. Matrix significantly predicted post-game single-item liking F (1, 108) = 4.52, p = .036, η_p^2 = .04. All other effects in all tested models were nonsignificant (all p s > .133). Stakes effects were all such that higher scores were in the risky condition, and the matrix effect on liking was such that the 9/1 (vs. 7/3) matrix had higher scores. The stakes results (effects on two of the four trust operationalizations and the closeness measure) suggest that forecasters might be able to discern the difference in the impact stakes can have on trust, but not consistently or to the same degree of impact that the stakes manipulation actually had on experiencers. The matrix effect on (one of two operationalizations of) liking could point to forecasters believing that the 9/1 matrix would promote liking, but given that this was the only significant matrix effect (and the complete lack of moderation by matrix, see Appendix D) it is difficult to draw strong conclusions based on this effect.

Appendix I: Study-Specific Effects on Key Variables

The only notable difference between experiential studies is that in the risky (vs. safe) condition participants in Study 2 report higher confidence in a typical undergraduate cooperating in Round One ($b = 0.64, p = .011, d = 0.62$), but the other studies also trend in the same direction.

Table II: Effect of Risky (vs. Safe) PD Game on Key Variables for Experiencers

Dependent variables	Study 1		Study 2		Study 3		Mega	
Cooperation	<i>t</i>	<i>d</i>	<i>t</i>	<i>d</i>	<i>t</i>	<i>d</i>	<i>t</i>	<i>d</i>
Confidence in partner R1 co-op.	0.50	0.11	0.84	0.20	0.26	-0.05	0.38	0.04
Confidence in undergrad R1 co-op.	1.44	0.31	2.62*	0.62	0.34	0.06	2.15*	0.25
Probability partner cooperates 100%	0.38	-0.08	0.60	0.14	0.45	-0.08	0.27	-0.03
Own cooperation rate	3.90***	0.82	3.88***	0.95	5.92***	1.04	7.54***	0.89
Reactions to PD game								
Cared about partner's choices	6.03***	1.30	1.12	0.26	7.21***	1.28	8.68***	1.01
Tempted to defect	1.37	-0.29	2.65*	-0.63	3.60***	-0.66	4.53***	-0.54
Worried partner would defect	0.10	0.02	1.61	-0.37	0.67	-0.13	1.17	-0.14
Gratitude and relief	—	—	7.56***	1.75	8.96***	1.63	12.00***	1.70
Trust								
Trust (single-item)	7.26***	1.56	5.54***	1.32	8.60***	1.58	12.30***	1.46
Reliance (single-item)	5.77***	1.23	4.33***	1.05	6.28***	1.15	9.88***	1.17
Trust (PD game impact)	9.29***	1.98	5.00***	1.15	9.59***	1.72	14.09***	1.64
Trust (perceived change)	8.41***	1.80	7.11***	1.76	12.09***	2.23	16.30***	1.96
Trust (slider)	7.33***	1.58	4.73***	1.13	6.07***	1.09	10.56***	1.24
Liking								
Single-item liking	3.79***	0.80	3.04*	0.75	4.49***	0.82	6.60***	0.79
More pleasant (perceived change)	9.15***	1.95	7.10***	1.78	13.21***	2.49	17.32***	2.10
Liking (slider)	2.04*	0.44	3.05**	0.77	4.88***	0.88	6.09***	0.72
Relationship								
Closeness (PD game impact)	6.71***	1.43	3.73***	0.86	8.37***	1.49	10.92***	1.26
Closer (perceived change)	6.68***	1.44	4.23***	1.03	10.19***	1.84	12.55***	1.49
Closeness (slider)	4.87***	1.05	3.63***	0.84	7.29***	1.30	9.28***	1.09
Stronger (perceived change)	7.16***	1.53	4.44***	1.08	8.08***	1.48	11.67***	1.39
Better (perceived change)	7.84***	1.67	5.15***	1.27	9.64***	1.80	13.46***	1.62

Note: Boldface indicates significant effects. PD = prisoner's dilemma, R1 co-op = Round One cooperation. Post-game sliders covary for pre-game sliders.

*** $p < .001$. ** $p < .01$. * $p < .05$. † $p < .1$.

Across forecasters studies we see very little variation, however looking at the results study-by-study shows us the mega-analyzed significant stakes effect on caring about partner's choices is being largely driven by Study 5 ($b= 0.55, p = .008, d = 0.31$). Additionally, in Study 4 participants in the risky (vs. safe) condition predicted less temptation to defect ($b= -0.81, p = .032, d = -0.42$), while in Study 5 they predicted more temptation to defect ($b= 0.51, p = .040, d = 0.23$).

Table I2: Effect of Risky (vs. Safe) PD Game on Key Variables for Forecasters

Dependent variables	Study 4		Study 5		Mega	
	<i>t</i>	<i>d</i>	<i>t</i>	<i>d</i>	<i>t</i>	<i>d</i>
Cooperation						
Confidence in partner R1 co-op. ^a	1.28	0.25	0.65	0.07	1.19	0.12
Confidence in undergrad R1 co-op.	1.28	0.25	—	—	—	—
Probability partner cooperates 100%	0.25	0.05	0.21	-0.02	0.04	0.00
Predicted coop rate (self)	0.48	0.09	1.59	-0.18	1.10	-0.11
Predicted coop rate (partner)	0.05	0.01	1.70 [†]	-0.19	1.40	-0.14
Reactions to PD game						
Cared about partner's choices	1.28	0.25	2.69**	0.31	2.84**	0.28
Tempted to defect	2.17*	-0.42	2.06*	0.23	0.77	0.08
Worried partner would defect	0.42	-0.08	0.59	0.07	0.23	0.02
Trust						
Trust (single-item)	0.49	0.11	0.43	0.05	0.74	0.07
Reliance (single-item)	0.02	0.01	0.36	-0.04	0.08	-0.01
Trust (PD game impact)	0.51	-0.12	1.13	0.13	0.47	0.05
Trust (perceived change)	0.21	0.04	1.15	0.13	1.18	0.11
Liking						
Liking (single-item)	0.36	0.08	0.47	0.05	0.67	0.07
More pleasant (perceived change)	0.07	-0.01	1.57	0.18	1.39	0.14
Relationship						
Closeness (PD game impact)	0.86	-0.20	1.05	0.12	0.16	0.02
Closer (perceived change)	0.52	0.10	1.39	0.16	1.48	0.14
Stronger (perceived change)	0.08	-0.02	0.37	0.04	0.33	0.03
Better (perceived change)	0.23	-0.05	0.26	0.03	0.19	0.02

Note. Boldface indicates significant effects. PD = prisoner's dilemma, R1 co-op = Round One cooperation. ^a Study 4 described their partner as a typical undergraduate, and so were only asked to predict what their partner would do in the first round.

[†] $p < .1$. * $p < .05$. ** $p < .01$. *** $p < .001$.

Appendix J: Nesting Structure of the Data

There was substantial dyad-level clustering for most primary variables (e.g., all trust operationalizations ICCs > .20, see Table J1), so we nested our analyses by dyad.

Table J1: Intraclass Coefficients and Descriptives for Experiencers' Focal Variables

Dependent Measures	Dyadic ICC
Cooperation	
Confidence in Round One partner cooperation	.23
Confidence in Round One undergrad cooperation	.12
Probability partner cooperates 100%	.01
Cooperation rate	.20
PD game reactions	
Cared about partners choices	.18
Tempted to defect	.17
Worried partner would defect	.10
Gratitude and relief ^a	.41
Trust-related	
Trust (single-item)	.33
Trust (perceived change)	.38
Trust (PD game impact)	.28
Reliance (single-item)	.20
Relationship-related	
Liking (single-item)	.27
More pleasant (perceived change)	.43
More close (perceived change)	.32
Closeness (PD game impact)	.44
Stronger (perceived change)	.32
Better (perceived change)	.35
Negotiation-related ^b	
Trust	.38
Trust (perceived change)	.26
Satisfaction	.14
Partner treatment	.21
Relationship satisfaction	.36
Engagement	.09
Info sharing	.58
Comfort	.18
Understanding	.23
Partner warmth and competence	.21
Cultural differences	.08
More pleasant (perceived change)	.19
Closer (perceived change)	.20
Stronger (perceived change)	.13
Better (perceived change)	.23

Note. ^a Gratitude and relief questions were included only in Studies 2 and 3. ^b Negotiation variables were only included in Studies 2 and 3.

Appendix K: Correlational Analyses for Key Variables

For experiencers we see substantial positive correlation between all trust operationalizations and generally moderate-high positive correlations between trust, liking, and closeness in both conditions. In the risky (vs. safe) condition, one's cooperation is moderately correlated with their partner's, and partner cooperation is positively correlated with trust, liking in only the risky condition. In the safe condition, own cooperation is moderately positively correlated with trust, liking and closeness. For all correlations, see Table K1.

Table K1: Bivariate Correlations for Experiencers

	1	2	3	4	5	6	7	8
(a) Primary Variables								
1. Trust (single-item)	—	.72**	.52**	.58**	.71**	.52**	.11	.18**
2. Reliance (single-item)	.66**	—	.43**	.49**	.69**	.48**	.04	.13*
3. Trust (PD game impact)	.27**	.27**	—	.49**	.42**	.70**	.07	.09
4. Trust (perceived change)	.42**	.31**	.29**	—	.57**	.44**	.15*	.18**
5. Liking (single-item)	.62**	.55**	.14*	.30**	—	.47**	.06	.15*
6. Closeness (PD game impact)	.28**	.28**	.74**	.27**	.15**	—	.05	.11
7. Own cooperation mean	.23**	.14*	.07	.22**	.16**	.02	—	.31**
8. Partner cooperation mean	.02	.08	-.07	.10	.08	-.01	.09	—
(b) Secondary Variables (Risky)								
10. Probability partner always coops	.09	.09	-.03	.06	.10	-.02	.12*	.05
11. Confidence in R1 partner coop	.20**	.09	.15*	.11	.18**	.12*	.18**	.09
12. Confidence in R1 undergrad coop	.05	.01	.07	-.03	.06	.04	.04	-.04
13. General trust	.24**	.21**	.05	.17**	.24**	.12*	.04	-.10
14. Social desirability	-.01	.03	-.04	.00	-.01	.02	.06	.03
15. Socio-economic status	-.05	-.02	.02	-.10	-.02	.09	-.05	-.06
16. Social-value orientation	.05	.03	.02	-.01	.08	-.01	.13*	-.04
(c) Secondary Variables (Safe)								
10. Probability partner fully coops	.29**	.22**	.21**	.17**	.24**	.23**	.24**	-.03
11. Confidence in R1 partner coop	.34**	.21**	.16**	.15*	.32**	.19**	.24**	.00
12. Confidence in R1 undergrad coop	.15*	.09	.05	.02	.15*	.01	.07	-.01
13. General trust	.20**	.12	-.13*	.06	.12*	-.02	.08	.10
14. Social desirability	.06	.07	-.09	.10	.10	-.05	.03	-.09
15. Socio-economic status	.14*	-.01	.10	.02	.03	.09	.03	.01
16. Social-value orientation	.16**	.16**	.03	.08	.18**	.10	.16**	-.03

Note. Boldface indicates significant effects. Top half of (a) is the bivariate correlations for the focal variables in the risky condition. Social-value orientation was coded as 0 = proself, 1 = prosocial.

Turning to forecasters, we see similar patterns for the primary outcomes, moderate-strong correlations across trust and liking in both conditions. PD game impact on closeness correlated consistently with trust in the risky condition (but less consistently in the safe condition). Own cooperation correlated with partner cooperation extremely highly in both conditions. Cooperation correlated with trust in both conditions (for all correlations see Table K2).

Table K2: Bivariate Correlations for Forecasters

	1	2	3	4	5	6	7	8
(a) Primary Variables								
1.Trust (single-item)	—	.72**	.26**	.57**	.77**	.16*	.21**	.31**
2.Reliance (single-item)	.73**	—	.19**	.51**	.72**	.14	.18*	.30**
3.Trust (PD game impact)	.18**	.18**	—	.31**	.19*	.76**	.20**	.24**
4. Trust (perceived change)	.53**	.58**	.22**	—	.54**	.18*	.22**	.32**
5.Liking (single-item)	.76**	.75**	.19**	.48**	—	.14	.14	.26**
6. Closeness (PD game impact)	.06	.09	.72**	.08	.14*	—	.09	.14
7. Own cooperation mean	.24**	.09	.00	.15*	.20**	.02	—	.80**
8. Partner cooperation mean	.32**	.19**	.06	.20**	.22**	.07	.79**	—
(b) Secondary Variables (Risky)								
10. Probability partner always coops	.24**	.25**	.20**	.32**	.25**	.14	.44**	.51**
11. Confidence in R01 partner coop	.23**	.26**	.20**	.17*	.20**	.13	.34**	.38**
12. Confidence in R01 undergrad coop	.30	.18	-.08	-.20	.21	-.10	.30*	.42**
13. General trust	.31**	.33**	.12	.30**	.32**	.05	.14*	.21**
14. Social desirability	.10	.09	-.08	.10	.11	-.15*	.16*	.15*
15. Socio-economic status	.18*	.12	.06	.10	.10	.04	.07	.13
16. SVO	-.09	-.14	.08	.15*	-.02	-.03	.24**	.14*
(c) Secondary Variables (Safe)								
10. Probability partner always coops	.37**	.26**	.09	.20**	.30**	.08	.46**	.51**
11. Confidence in R01 partner coop	.24**	.23**	.08	.17*	.18*	.14*	.32**	.43**
12. Confidence in R01 undergrad coop	.31*	.28*	.02	.22	.34*	.03	.47**	.54**
13. General trust	.36**	.34**	.11	.24**	.35**	.12	.23**	.30**
14. Social desirability	.11	.03	-.08	.04	.06	-.10	.07	.02
15. Socio-economic status	-.05	-.03	-.06	-.01	.00	-.06	-.03	-.11
16. SVO	.06	.04	.16*	.04	.05	.19**	.22**	.13

Note. Boldface indicates significant effects. Top half of (a) is the bivariate correlations for the focal variables in the risky condition. Social-value orientation was coded as 0 = prosocial, 1 = prosocial. Own cooperation and partner cooperation were predicted by forecasters.

Appendix L: Tests of Closeness Moderation

To confirm that the effects of the PD game are not entirely due to closeness, we tested moderation of pre-game closeness scores on stakes effects on key outcomes. There was only one significant interaction, such that experiencers with higher pre-game closeness report higher perceived change in closeness in the risky (vs. safe) condition (see Table L1 for all closeness moderation tests).

Table L1: Closeness Moderation of Stakes Effects for Experiencers

Variable	Experiencers				
	Stakes effect				Stakes × Closeness
	Low closeness		High closeness		
	<i>t</i>	<i>d</i>	<i>t</i>	<i>d</i>	<i>t</i>
Cooperation					
Confidence in partner R1 co-op	0.63	0.09	0.42	-0.06	-0.75
Confidence in undergrad R1 co-op	0.70	0.11	2.23*	0.37	1.10
Probability partner cooperates 100%	0.81	0.11	1.32	-0.19	-1.51
Own cooperation rate	6.03***	0.80	4.72***	0.65	-0.91
Partner cooperation rate	4.92***	0.65	5.82***	0.82	0.68
PD game reactions					
Cared about partners choices	5.20***	0.74	6.86***	0.96	1.23
Tempted to defect	3.62***	-0.49	2.75**	-0.41	0.61
Worried partner would defect	0.81	-0.12	0.86	-0.13	-0.04
Gratitude and relief	8.87***	1.42	8.20***	1.55	0.18
Trust					
Trust (single-item)	10.30***	1.40	8.18***	1.21	-1.47
Reliance (single-item)	7.57***	1.08	6.82***	0.99	-0.49
Trust (PD-game impact)	10.00***	1.45	9.73***	1.39	-0.14
Trust (perceived change)	10.73***	1.53	12.40***	1.78	1.25
Trust (post-game slider)	7.71***	1.09	7.26***	1.12	-0.27
Liking					
Liking (single-item)	4.73***	0.71	4.91***	0.72	0.16
More pleasant (perceived change)	11.38***	1.69	13.37***	1.91	1.50
Liking (post-game slider)	3.69***	0.55	4.50***	0.69	0.59
Relationship					
Closeness (PD-game impact)	6.86***	0.99	8.91***	1.27	1.52
Closer (perceived change)	7.70***	1.08	10.50***	1.51	2.07*
Closeness (post-game slider)	5.62***	0.80	7.73***	1.09	1.56
IOS (post-game)	5.39***	0.87	7.62***	1.13	1.60
Stronger (perceived change)	7.48***	1.16	9.30***	1.37	1.38
Better (perceived change)	8.81***	1.31	10.27***	1.48	1.12
Negotiation					
Info sharing	1.36	0.21	1.46	0.27	0.19
Trust during negotiation	2.03*	0.32	2.45*	0.45	0.50
Perceived change in trust	0.75	0.12	0.85	0.15	0.14
More pleasant (perceived change)	0.57	0.09	1.34	0.24	0.63

Note. Boldface indicates significant effects. R1 = Round One. IOS = inclusion of other in self. ^a Confidence in typical undergraduate Round One cooperation was not measured in Study 5.

Slider measures, IOS, and negotiation-related variables were only collected for experiencers. *** $p < .001$. ** $p < .01$. * $p < .05$. † $p < .1$.