When Social Constraints Increase Trust: Considering Causal Attributions as a Source of Treatment Effect Heterogeneity

Article · September 2016

DOI: 10.1177/2378023116667360

1 author:

Blaine G. Robbins
New York University Abu Dhabi
35 PUBLICATIONS **125** CITATIONS

SOCIUS
Original Article
Socius: Sociological Research for a Dynamic World
Volume 2: 1–11
© The Author(s) 2016
DOI: 10.1177/2378023116667360
srd.sagepub.com
SAGE

When Social Constraints Increase Trust: Considering Causal Attributions as a Source of Treatment Effect Heterogeneity

Blaine G. Robbins[1,2]

Abstract

The degree to which social constraints promote or undermine trust remains unknown. One classic perspective suggests that trust blossoms in the presence of social constraints, while another influential school of thought proposes that social constraints wither trust. The author integrates both traditions and proposes a model whereby social constraints increase trust, but only to the extent that individuals attribute another's perceived trustworthiness to the situation. As individuals increasingly attribute another's perceived trustworthiness to dispositional factors, the positive effect of social constraints on trust declines and approaches zero. The author addresses this debate and tests the model by designing two novel survey experiments of simulated car repair and group project scenarios. Findings from two large crowdsourced samples support the model. Implications for existing theory and future research are discussed.

Keywords

attributions, social constraints, trust, trustworthiness, survey experiment

Does trust blossom or wither in the presence of social constraints? One influential school of thought common to economics and political science suggests that social constraints are an effective tool for establishing trust between two parties (Farrell 2009; Greif 2006; Knight 2001; North 1990; Rothstein and Stolle 2008; Sztompka 1999). Another equally influential growing body of work found in sociology and social psychology suggests the opposite: social constraints increase trust in the short run but ultimately undermine trust in the long run (Irwin, Mulder, and Simpson 2014; Malhotra and Murnighan 2002; Mulder et al. 2006; Simpson and Eriksson 2009). I contribute to this debate by integrating elements from both perspectives and by arguing that social constraints do indeed increase trust, but only to the extent that individuals attribute another's perceived trustworthiness to the situation (Gilbert and Malone 1995). If actors draw situational attributions and perceive another's trustworthiness as extrinsically motivated, then social constraints have strong positive effects on trust. If, on the other hand, actors draw dispositional attributions and perceive another's trustworthiness as intrinsically motivated, then social constraints have null to weak positive effects on trust.

In administering two survey experiments to Amazon.com Mechanical Turk (MTurk) workers (n = 1,388 and n = 1,419), I show that the causal relation between social constraints and trust is moderated by causal attributions: social constraints increase trust for those who gravitate toward external attributions, but not for those who tend toward internal attributions, while internal attributions produce greater trust than external attributions, even in the presence of social constraints. These findings provide support for my integrated model's core claim and have implications for future research and theory.

Social Constraints, Causal Attributions, and Trust

Trust is a belief about another person's trustworthiness with respect to a particular matter at hand that emerges under conditions of unknown outcomes, where trustworthiness is the capability (competence and ability) and commitment (exertion and motivation) of a trustee (Hardin 2002; Robbins forthcoming-b). Trust, in other words, is a dyadic (Wilson

[1]New York University Abu Dhabi, Abu Dhabi, United Arab Emirates
[2]University of California, Berkeley, CA, USA

Corresponding Author:
Blaine G. Robbins, New York University Abu Dhabi, Saadiyat Island, Building A5, Abu Dhabi, United Arab Emirates.
Email: bgr3@nyu.edu

and Eckel 2011) and relational concept (Cook, Hardin, and Levi 2005) in which actor A (the truster) trusts actor B (the trustee) when actor A believes that actor B is capable and committed to perform matter Y (what A wants B to do) under conditions of unknown outcomes.[1] According to this definition, trustworthiness begets trust, yet which forms of trustworthiness are necessary and or sufficient to produce trust is an outstanding question.

Drawing on Hobbesian notions of social order, one school of thought suggests that social constraints—interventions external to or beyond an exchange relationship that influence behavior by altering the costs and benefits of action—are sufficient for trust to develop (Farrell 2009; Greif 2006; Knight 2001; North 1990; Rothstein and Stolle 2008). Binding contracts, legal regulations, and social norms are examples of exogenous motivators that foster trustworthiness: the breach of contract, the force of law, and the promise of collective shaming respectively motivate individuals to act in the interests of others. Under these conditions, the costs and benefits associated with social constraints align the interests of both parties and compel a trustee, out of self-interest, to act trustworthily toward a truster. With this *deterrence-based* view of trust, trustworthiness is realized when the interests of one party encapsulate the other (Farrell 2009), and as a result, any social constraint external to an exchange relationship is sufficient to produce trustworthiness and, hence, trust.

An alternative school of thought holds that social constraints undermine cooperation and trust (Titmuss 1970). Scholars in this area suggest that when an actor is intrinsically motivated to perform a given act but is subject to an external reward or punishment, the actor will attribute his or her behavior to the extrinsic device rather than to intrinsic desires (Deci, Koestner, and Ryan 1999). External motivators in this instance replace, or *crowd out*, internal motivations and an actor's intrinsic desire to perform a given act decreases. The implication being that intrinsically motivated cooperation dissipates through time in the presence of social constraints.

Recent work suggests that the scope of the crowding-out effect extends beyond one's own intrinsic motivations to beliefs about the intrinsic motivations of others. Mulder et al. (2006) and colleagues (Irwin et al. 2014) argued that social constraints serve as signals of distrust and undermine trust when used (see also Chen, Pillutla, and Yao 2009; Kuwabara 2015; Malhotra and Murnighan 2002; Simpson and Eriksson 2009). To test this proposition, the authors used a method known as *removing the sanction* (Deci 1971). The logic behind this method suggests that levels of trust should decline for treatment groups when sanctions are introduced and then removed. Control groups sans sanctions should not experience similar declines in trust. The argument is that social constraints simultaneously increase the belief that trusted actors are externally motivated to cooperate but decrease the belief that trusted actors are internally motivated to cooperate. In other words, social constraints act as countervailing forces that promote and undermine trust via causal attributions.

In an effort at synthesis and integration, I argue that social constraints increase trust but that causal attributions moderate—not mediate as suggested by the crowding-out tradition—the relation between social constraints and trust. To put it differently, social constraints increase trust but the size and magnitude of this effect is conditional on whether the trustworthiness of others is perceived as intrinsically or extrinsically motivated. To make this claim, I draw on the attribution bias literature in social psychology, which finds that peoples' attributions regarding the causes of their own and others' behaviors do not always mirror reality (Jones and Harris 1967; Ross 1977). What the attribution bias literature shows is that some individuals correctly estimate the intrinsic and extrinsic motivations behind other peoples' behavior, while some individuals do not (Gilbert and Malone 1995).

Drawing on the arguments above, I take the next logical step and contend that causal attributions influence not only *explanations* of observed behavior but also *expectations* of future behavior, even in the absence of prior behavior. I make such claims because the mechanisms accounting for attribution biases such as the fundamental attribution error (FAE) and correspondence bias (CB) are activated when expectations about a person's trustworthiness are formed (Gilbert and Malone 1995; Jones and Harris 1967; Ross 1977).[2] First, a person may *lack awareness* of the causal role a situational force might play in the prior and future behavior of others. Second, a person may be aware of the causal role a situational force might play but hold *unrealistic expectations* about how the situation affects another's prior and future behavior. Taken together, *lack of awareness* and *unrealistic*

[1]Other common definitions of trust in the literature underscore trusting people in general (e.g., strangers, fellow citizens) and for all matters, what some refer to as generalized trust, social trust, or general social trust (Putnam 2000; Rothstein 2000; Uslaner 2002; see Nannestad 2008 for a review). The definition of trust used here, in contrast, centers on trust in specific people for particular matters (see Robbins forthcoming-b).

[2]The social psychological dynamics behind the FAE and other forms of CB are central to my integrated model. FAE is the tendency to overestimate the effect of disposition (or internal psychology) and underestimate the effect of the situation (or external factors) in explaining social behavior, while CB is the tendency to draw dispositional inferences from behavior regardless of situational factors. For instance, if you provide me with poor service at a restaurant, I might attribute your poor service to laziness or carelessness—dispositional factors—whereas, in reality, a death in the family—the situation—influenced your poor service. If I, as a customer, was aware of the death in your family but still attributed your poor service to laziness and carelessness, FAE would be at work. If I did not know about the death in your family and attributed your poor service to laziness and carelessness, CB would be at work.

expectations should produce expectations of trustworthiness that underestimate the causal role of situational forces.

I thus expect social constraints to increase trust, with stronger positive effects when a person's trustworthiness is perceived as externally motivated and weaker positive effects when a person's trustworthiness is perceived as internally motivated, where the latter is driven by *lack of awareness* and *unrealistic expectations*.[3] I also expect greater trust among actors who underestimate the effects of social constraints (and draw dispositional inferences regarding another's trustworthiness) than actors who attribute the effects of social constraints to the situation because perceived internal motivations produce greater trust than perceived external motivations (Mayer, Davis, and Schoorman 1995; Yamagishi and Yamagishi 1994). In short, social constraints do not necessarily undermine or crowd-out trust. Instead, the deterrence-based effect of social constraints on trust varies to the extent that one draws dispositional or situational attributions of another's perceived trustworthiness.[4]

See Figure 1 for causal diagrams illustrating the deterrence-based view, the crowding-out perspective, and my integrated model

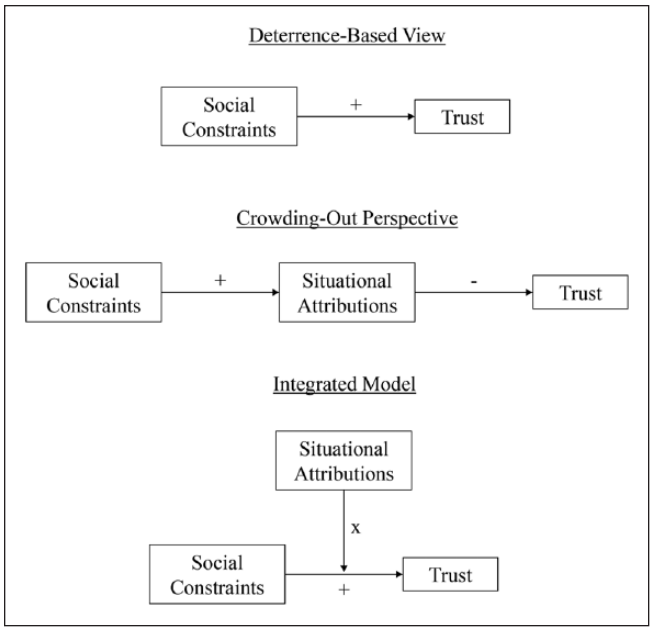

Figure 1. Models of trust from the deterrence-based view, the crowding-out perspective, and an integrated model.

Method

Design, Participants, and Procedure

To examine the causal relation between perceived trustworthiness and trust, I use a factorial survey experiment design (Auspurg and Hinz 2015; Hainmueller, Hopkins, and Yamamoto 2014; Jasso 2006; Rossi and Knock 1982). This approach presents respondents with hypothetical scenarios containing situational and relational conditions of theoretical importance. With this method, a researcher systematically manipulates features of a social context that theoretically influence judgment-making processes of interest. For the present study, I have created two hypothetical scenarios, a car repair scenario (study 1a) and a group project scenario (study 1b), each consisting of 10 dimensions, in which subjects are asked the extent to which they trust a hypothetical auto mechanic or group project member. In creating the vignette scenarios and dimensions, I was guided by the trust literature, my integrated model, and a pilot study. The two scenarios were developed to explore the robustness of findings under different conditions, such as the principal-agent problem (car repair scenario) versus the collective action problem (group project scenario), rather than test hypothesized differences in parameter estimates across scenarios.

Each hypothetical scenario features a 5 (age: 20, 30, 40, 50, or 60 years) × 4 (race: white, black, Hispanic, or Asian) × 2 (gender: male or female) × 2 (reputation: no reputation or positive reputation) × 3 (halo: blank, bad used computer, or good used computer) × 2 (competence: blank or competent) × 2 (exertion: blank or hard-working) × 6 (perceived internal motivations: uncooperative, no prior interaction, prior interaction, encapsulated interests, goodwill, or virtuous dispositions) × 3 (contract: blank, nonbinding contract, or binding contract) × 3 (regulation: no regulations, nonmonetary regulations, or monetary regulations) multifactorial vignette design, which yields a factorial object universe of 51,840 ($2^4 \times 3^3 \times 4^1 \times 5^1 \times 6^1$) unique vignettes in which all possible combinations of dimensions were included in the factorial object universe. The dimensions for age, race, gender, reputation, halo, competence, exertion, and perceived internal motivations will be explored in other papers (e.g., Robbins forthcoming-a).

I administered a Web-based version of my survey experiments to Amazon.com MTurk workers during the fall of 2013. A total of 1,388 workers participated in study 1a (52 percent men, mean age = 32.61 years, *SD* = 11.51 years), while 1,419 workers participated in study 1b (44 percent men, mean age = 32.10 years, *SD* = 10.89 years). To be eligible, workers must have been legal adults residing in the United States with approval rates of 90 percent or above on previous tasks. No worker participated in both studies. The overall sample size was determined by a power analysis, and my data

[3]I expect similar processes regardless of culture: only the rate of attribution biases should vary from culture to culture (Morris and Peng 1994).

[4]Yamagishi and Yamagishi (1994) distinguished between expectations based on inferences about another's internal motivations (*trust*) and external motivations (*assurance*). Following the encapsulated-interests literature (Cook et al. 2005; Farrell 2009; Hardin 2002), I contend that social constraints, a form of external motivation, are sufficient to produce trust but that their effects depend on if and how they are perceived.

collection stopping rule for each study consisted of reaching a target sample size of 1,350 respondents who completed *and* submitted a human intelligence task to Amazon.com (see the Supplemental Materials online for more information).

After consenting to participate, workers were shown a coversheet asking respondents to imagine a hypothetical car repair or group project scenario. Participants were then quizzed on the respective scenarios and shown 10 vignettes randomly drawn with replacement from the vignette object universe of 51,840 unique vignettes. Although the levels of each dimension were randomized, the order of dimensions was fixed from vignette to vignette. After assessing the 10 vignettes, participants filled out a demographic questionnaire, were shown a debriefing statement, thanked for their participation, and then paid $2. The median time respondents participated in studies 1a and 1b was 18.12 and 18.87 minutes, respectively.

Measures

Vignette dimensions for social constraints. The common types of social constraints investigated in the crowding-out literature include binding and nonbinding contracts (Malhotra and Murnighan 2002), monetary punishments and rewards (Irwin et al. 2014; Mulder et al. 2006), and moral incentives (Chen et al. 2009). To operationalize the constellation of social constraints, I use two separate dimensions (*contract* and *regulation*), each consisting of three levels.

Drawing on Malhotra and Murnighan (2002), levels for the *contract* dimension were operationalized as *no contract, nonbinding contract*, and *binding contract*. Nonbinding contracts, such as handshakes and verbal promises, facilitate exchange sans written agreements; binding contracts, on the other hand, encourage exchange with written agreements enforceable by organizational rules and laws. The contract dimension thus operationalizes social constraints as centralized and decentralized controls that emerge from and apply to specific exchange relations (via nonbinding verbal promises or binding agreements). My expectation is that nonbinding contracts will have weak positive effects on trust regardless of attribution type (Malhotra and Murnighan 2002), be it dispositional or situational, because nonbinding contracts by definition lack external controls that restrict and incentivize behavior. I expect binding contacts to follow the dynamics outlined in my integrated model: strong positive effects on trust for situational attributions and weak positive effects on trust for dispositional attributions.

Levels for the *regulation* dimension included the following: *no regulation, nonmonetary regulation*, and *monetary regulation*. Drawing on Chen et al. (2009), I operationalize the nonmonetary regulation level as a centralized control that polices multiple exchange relations with moral incentives (e.g., mandatory business ethics classes for transgressions), while the monetary regulation level centers on financial incentives (e.g., fines). I expect both forms of regulation to suffer from attribution biases and generate interaction effects outlined in my integrated model.

Endogenous variables. At the bottom of each vignette, participants were shown two questions. One question assessed participants' trust, and the other assessed participants' causal attributions. The trust question was structured as an 11-point bipolar item and asked participants the following (group project elements in brackets): "Given the conditions above, to what extent do you trust the auto mechanic [student] to provide justifiable and quality auto repairs [to complete the assigned data analysis task]?" Response options ranged from "complete distrust" (0 value) through "neither trust nor distrust" (50 value) to "complete trust" (100 value), with a "don't know" option at the end of the item (study 1a: M = 6.51, SD = 2.68, minimum = 0, maximum = 10; study 1b: M = 7.49, SD = 2.30, minimum = 0, maximum = 10).

The causal attributions question was also structured as an 11-point bipolar item and asked participants about their reported level of causal attributions (group project elements in brackets): "Was the value you provided above primarily influenced by characteristics of the situation, primarily influenced by characteristics of the mechanic [student], or influenced by characteristics of both (if you think both contributed equally, mark 50 on the scale)?" Response options ranged from "characteristics of the situation" (0 value) through "characteristics of both" (50 value) to "characteristics of the mechanic [student]" (100 value), with a "don't know" option at the end of the item (study 1a: M = 5.85, SD = 2.43, minimum = 0, maximum = 10; study 1b: M = 6.40, SD = 2.36, minimum = 0, maximum = 10). In other words, participants were shown two evaluation tasks that ranged from 0 to 100 (increasing by increments of 10), which were recorded as 11-point bipolar variables ranging from 0 to 10 (increasing by increments of 1). Trust and causal attributions responses consisting of "don't know" (<0.25 percent) were excluded from the analysis.

Individual-level covariates. Because respondents vary with respect to when, where, and how they participated in studies 1a and 1b, I control for a number of individual-level covariates, W_j, intended to reduce model noise and address nonindependence of observations (among others) that can arise from such unsystematic variation. First, it is possible for MTurk workers to participate in study 1a or 1b from the same Internet protocol (IP) address. In this case, data are likely correlated and standard errors downwardly biased, because clustered observations that violate the "stable unit treatment value assumption" contain less unique information.[5] To address this issue, I include dummy variables in which the referent category represents all IP addresses with a single partial or

[5]This issue can take multiple forms, such as (1) the same worker with different MTurk worker identification numbers participating in a single study from the same IP address (MTurk workers can create unique worker identification numbers for any number of credit cards they own), (2) cohabitants with different MTurk worker identification numbers participating in a single study from the same IP address, and (3) different workers from the same masked IP address.

complete experiment and the indicator categories represent a vector of IP addresses with multiple partial or complete experiments (study 1a: 11 percent of participants from similar IP addresses; study 1b: 7 percent of participants from similar IP addresses). Second, my models contain a binary variable in which the referent category represents complete experiments and the indicator category represents partial experiments (study 1a: 2 percent of participants were partials; study 1b: 4 percent of participants were partials). This is done to account for problems of attrition.

Third, as a manipulation check and to reduce overall model noise (Berinsky, Margolis, and Sances 2014), I include two binary variables for each true-or-false screener question administered directly after the coversheet (and prior to the 10 vignettes), in which the referent category represents an incorrect answer and the indicator category represents a correct answer (study 1a: screener 1 = 96 percent correct, screener 2 = 98 percent correct; study 1b: screener 1 = 95 percent correct, screener 2 = 94 percent correct). Fourth, I further reduce model noise attributed to issues of attention by controlling for length of participation in a study (natural logarithm of time in minutes) (study 1a: $M = 2.90$, $SD = 0.39$, minimum = 0.31, maximum = 4.70; study 1b: $M = 2.94$, $SD = 0.42$, minimum = 0.41, maximum = 5.47). Fifth, to account for history effects, I include a binary variable in which the referent category represents the first day of data collection and the indicator category represents the second day of data collection (data collection was completed in two days for study 1a and study 1b).

Analytic Strategy

My factorial research design yields panel data in which i vignettes ($i = 1, \ldots, 10$) are nested within j individuals ($j = 1, \ldots, J$). As a result, I estimated two-level correlated random-effects models with moderation in which lower level moderation (i.e., causal attributions) of lower level effects (i.e., social constraints and trust) takes place (Mundlak 1978; Wooldridge 2010).

The level 1 (or within-level) model takes the following form:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \delta_i V_i + e_{ij}, \quad (1)$$

where Y_{ij} is a continuous measure of trust in the ith vignette for the jth individual, β_{0j} is a random intercept term capturing unobserved heterogeneity varying across individuals but not vignettes, β_{1j} is a nonrandom slope for X_{ij} (which is a vector of vignette dimensions treated as dummy variables that vary across both individuals and vignettes), δ_i is a nonrandom slope for V_i (which is the ith vignette treated as $i - 1$ dummy variables), and e_{ij} is a disturbance term that varies over the population of vignettes (assumed normal, independent, and identically distributed).

I specify a level 2 model of between-individual variation in trust by modeling the random intercept, β_{0j}, from equation 1:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \bar{X}_j + \gamma_{02} W_j + u_{0j}, \quad (2)$$

where β_{0j} is a random intercept term capturing individual-level variation in trust, γ_{00} is an overall population intercept for trust, γ_{01} is a nonrandom slope for $\bar{X}_j$ (which is a vector of individual-specific means for X_{ij} that vary across individuals but not vignettes), γ_{02} is a nonrandom slope for W_j (which is a vector of individual-level covariates), and u_{0j} is a random disturbance term that varies over the population of individuals (assumed normal, independent, and identically distributed).

Because I randomly assigned levels of each dimension to vignettes, I can safely assume that X_{ij} are orthogonal to e_{ij}, the level 1 disturbance term. I cannot, however, safely assume that X_{ij} are orthogonal to u_{0j} even though I randomly assigned levels of each dimension to individuals: unbalanced distributions of levels of dimensions between individuals might correlate with u_{0j}. That is, individuals with greater (lower) proportions of certain levels of a dimension across the 10 vignettes may produce higher (lower) mean levels of trust via learning and fatigue effects for instance. If X_{ij} and u_{0j} covary, then X_{ij} conflate within- and between-individual components, yielding inconsistent but efficient estimates of β_{1j}. Hausman specification tests estimating fixed- and random-effects econometric models for panel data support this conclusion (study 1a: $\chi^2[22] = 66.03$, $p < .001$; study 1b: $\chi^2[22] = 56.72$, $p < .001$).

As a result, I include $\bar{X}_j$ in equation 2 to decompose (or deconflate) the variances of X_{ij} into within- and between-individual components, yielding unbiased and consistent estimates of β_{1j}. Because $\bar{X}_j \neq 0$, equation 2 allows β_{0j} to be correlated with $\bar{X}_j$, which makes β_{1j} a within-individual between-vignette estimator (or fixed effect) and γ_{01} represents the difference between the within- and between-individual effect (or the unique contextual effect). In short, the inclusion of $\bar{X}_j$ in equation 2 coupled with the orthogonality assumption of X_{ij} and e_{ij} implies that I can interpret β_{1j} as causal and unbiased.

To explore heterogeneous treatment effects of social constraints on trust by causal attributions (i.e., hypothesis 1), I included the following terms to equation 1: $\beta_{2j} A_{ij}$ and $\beta_{3j} X_{ij} A_{ij}$, where β_{2j} is a nonrandom slope for A_{ij} (which is a continuous measure of causal attributions) and β_{3j} is a nonrandom slope for $X_{ij} A_{ij}$ (which is a within-individual interaction between X_{ij} and A_{ij}). Thus,

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \beta_{2j} A_{ij} + \beta_{3j} X_{ij} A_{ij} + \delta_i V_i + e_{ij}. \quad (3)$$

To equation 2, I added $\gamma_{03} \bar{A}_j$ and $\gamma_{04} \overline{X_j A_j}$, where γ_{03} is a nonrandom slope for $\bar{A}_j$ (which are individual-specific means for A_{ij} that vary across individuals but not vignettes), and γ_{04} is a nonrandom slope for $\overline{X_j A_j}$ (which are individual-specific means for the interaction between X_{ij} and A_{ij}). Like $\bar{X}_j$, I include $\bar{A}_j$ and $\overline{X_j A_j}$ to decompose the effects of A_{ij} and $X_{ij} A_{ij}$, respectively, into within- and between-individual components (Schunck 2013). Thus,

Table 1. Two-level Correlated Random-effects Models of Trust with Moderation.

	Model 1	Model 2	Model 3	Model 4
	Car Repair	Car Repair	Group Project	Group Project
No contract (reference)				
Nonbinding contract	–0.04 (.04)	–0.06 (.11)	0.10** (.03)	0.15 (.11)
	–0.01	**–0.004**	**0.02**	**0.02**
Binding contract	0.61*** (.04)	1.06*** (.11)	0.37*** (.03)	0.71*** (.10)
	0.11	**0.11**	**0.08**	**0.08**
No regulation (reference)				
Nonmonetary regulation	0.56*** (.04)	1.00*** (.12)	0.23*** (.03)	0.42*** (.11)
	0.10	**0.10**	**0.05**	**0.05**
Monetary regulation	0.75*** (.04)	1.41*** (.11)	0.56*** (.03)	1.20*** (.11)
	0.13	**0.13**	**0.12**	**0.12**
Causal attributions		0.18*** (.02)		0.15*** (.02)
		0.09		**0.08**
Nonbinding × Attributions		0.01 (.02)		–0.01 (.02)
		0.002		**–0.004**
Binding × Attributions		–0.07*** (.02)		–0.05*** (.015)
		–0.03		**–0.02**
Nonmonetary × Attributions		–0.08*** (.02)		–0.03† (.02)
		–0.03		**–0.01**
Monetary × Attributions		–0.11*** (.02)		–0.10*** (.02)
		–0.05		**–0.05**
Constant	5.54*** (.51)	4.30*** (.69)	5.45*** (.50)	3.60*** (.89)
var(u_{0j})	0.69*** (.06)	0.67*** (.06)	0.70*** (.04)	0.63*** (.04)
var(e_{ij})	2.95*** (.07)	2.90*** (.07)	2.17*** (.06)	2.14*** (.05)
Other vignette dimensions	Yes	Yes	Yes	Yes
Vignette dummies	Yes	Yes	Yes	Yes
Individual-specific mean dimension	Yes	Yes	Yes	Yes
Individual-level covariates	Yes	Yes	Yes	Yes
Observations	13,733	13,733	14,019	14,019
Individuals	1,383	1,383	1,414	1,414

Note: Shown are unstandardized slopes (robust standard errors in parentheses), with standardized slopes in boldface type.
†$p < .10$, *$p < .05$, **$p < .01$, and ***$p < .001$ (two tailed).

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\bar{X}_j + \gamma_{02}W_j + \gamma_{03}\bar{A}_j + \gamma_{04}\overline{X_j A_j} + u_{0j}. \quad (4)$$

Further Details

Further details about my samples and procedures, Amazon.com MTurk, the exact wording of the vignette scenarios and endogenous variables, and model-based and design-based assumptions can be found in the Supplemental Materials online. See Robbins (2016) for a discussion of the research design's merits and strengths, especially in relation to behavioral economic games.

Results

Although not shown, null models for the car repair and group project scenarios yielded relatively small intraclass correlation coefficients (model 1 = .10, model 2 = .14), and statistical significance was achieved for the level 2 disturbance terms in both the car repair ($u_{0j} = 0.71$, $SE = 0.05$, $p < .001$) and group project scenarios ($u_{0j} = 0.73$, $SE = 0.05$, $p < .001$). This suggests that 90 percent and 86 percent of the variation in trust is accounted for by contextual and situational characteristics of the car repair and group project scenarios, respectively.

Table 1 presents results of the correlated random-effects models predicting trust in the car repair (models 1 and 2) and group project (models 3 and 4) scenarios. Models 1 and 3 provide main effects, while models 2 and 4 provide interaction effects. With respect to main effects, models 1 and 3 show that there was a statistically significant effect of contract on trust at the $p < .05$ level (model 1: $\chi^2[2] = 304.56$, $p < .001$; model 3: $\chi^2[2] = 131.39$, $p < .001$). Post hoc comparisons using the Wald test indicated that binding contracts were significantly different from and produced greater trust than nonbinding contracts (model 1: $\chi^2[2] = 252.61$, $p < .001$; model 3: $\chi^2[2] = 60.26$, $p < .001$). Table 1 also shows that the binding contract condition was significantly different from and produced greater trust than the no contract

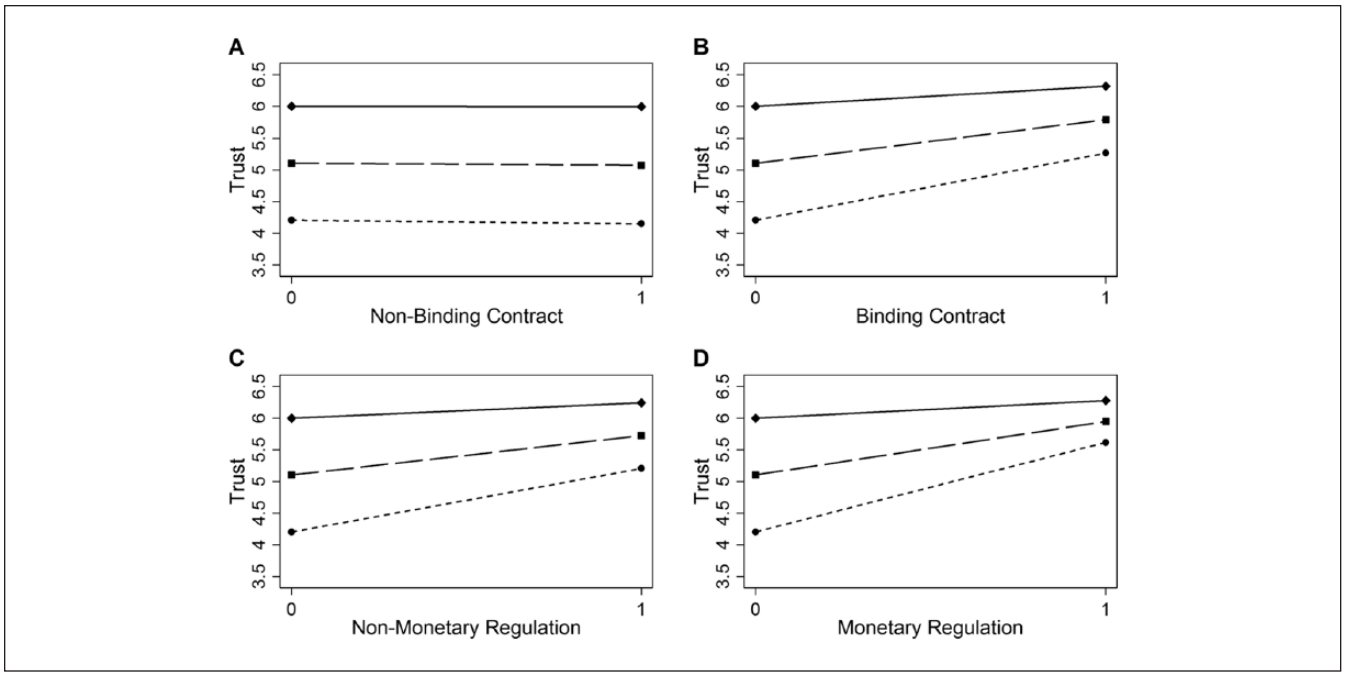

Figure 2. Relationship between social constraints and trust by causal attributions in the car repair scenario (Table 1, model 2).
Note: For panels A to D, solid black slopes indicate dispositional attributions (10), dashed black slopes indicate both dispositional and situational attributions (5), and dotted black slopes indicate situational attributions (0).

condition, but that the nonbinding contract condition was only significantly different from and produced greater trust than the no contract condition in model 3. Likewise, there was a statistically significant effect of regulation on trust at the $p < .05$ level (model 1: $\chi^2[2] = 368.39$, $p < .001$; model 3: $\chi^2[2] = 268.27$, $p < .001$). Post hoc comparisons using the Wald test indicated that monetary regulations were significantly different from and produced greater trust than nonmonetary regulations (model 1: $\chi^2[1] = 23.22$, $p < .001$; model 3: $\chi^2[1] = 100.03$, $p < .001$). Table 1 also shows that the nonmonetary regulation and monetary regulation conditions were significantly different from and produced greater trust than the no regulation condition.

With respect to interaction effects, models 2 and 4 support the dynamics predicted by my integrated model and show that there was a statistically significant interaction between contract and attributions at the $p < .05$ level (model 2: $\chi^2[2] = 24.70$, $p < .001$; model 4: $\chi^2[2] = 12.12$, $p < .01$), and regulation and attributions at the $p < .05$ level (model 2: $\chi^2[2] = 41.69$, $p < .001$; model 4: $\chi^2[2] = 39.12$, $p < .001$). Moreover, Table 1 indicates that binding contracts, nonmonetary regulations, and monetary regulations yielded significant positive effects on trust when situational attributions were drawn (i.e., zero on the causal attributions scale), but that these positive effects declined, to varying degrees depending on the type of social constraint, as causal attributions moved from situational to dispositional (i.e., from 0 to 10 on the causal attributions scale). See Figures 2 and 3 for an illustration of these dynamics.

Figures 4 and 5 graphically illustrate the marginal effects of social constraints on trust. The solid black line in each panel indicates how the slopes for social constraints change across causal attributions. Ninety-five percent confidence intervals indicate whether, for a given slope along the range of causal attributions, social constraints significantly affect trust at the $p < .05$ level. This occurs when the upper and lower bounds of a confidence interval are above (or below) the dashed zero line. For instance, in Figures 4, the slopes for monetary regulations in both scenarios were relatively steep when trustworthiness was attributed to situational causes (model 2: $b = 1.41$; model 4: $b = 1.20$). But as causal attributions moved from fully situational to fully dispositional, the slopes in both scenarios gradually decreased in size, eventually reaching $b = 0.28$ in model 2 and $b = 0.26$ in model 4. As expected, similar dynamics were observed in Figure 4 for binding contracts and nonmonetary regulations but not for nonbinding contracts.

Unlike Figure 4, the marginal effects in Figure 5 reveal additional meaningful information beyond what can be inferred from the statistically significant interaction terms in model 4. First, when attributions were described as both situational *and* dispositional the effect of nonbinding contracts on trust was statistically significant, albeit negligible in terms of effect size. In other words, for all values indicating fully situational or fully dispositional attributions (i.e., less than 4 or greater than 8, respectively), the presence of nonbinding contracts did not significantly affect trust. Second, the positive effect of nonmonetary

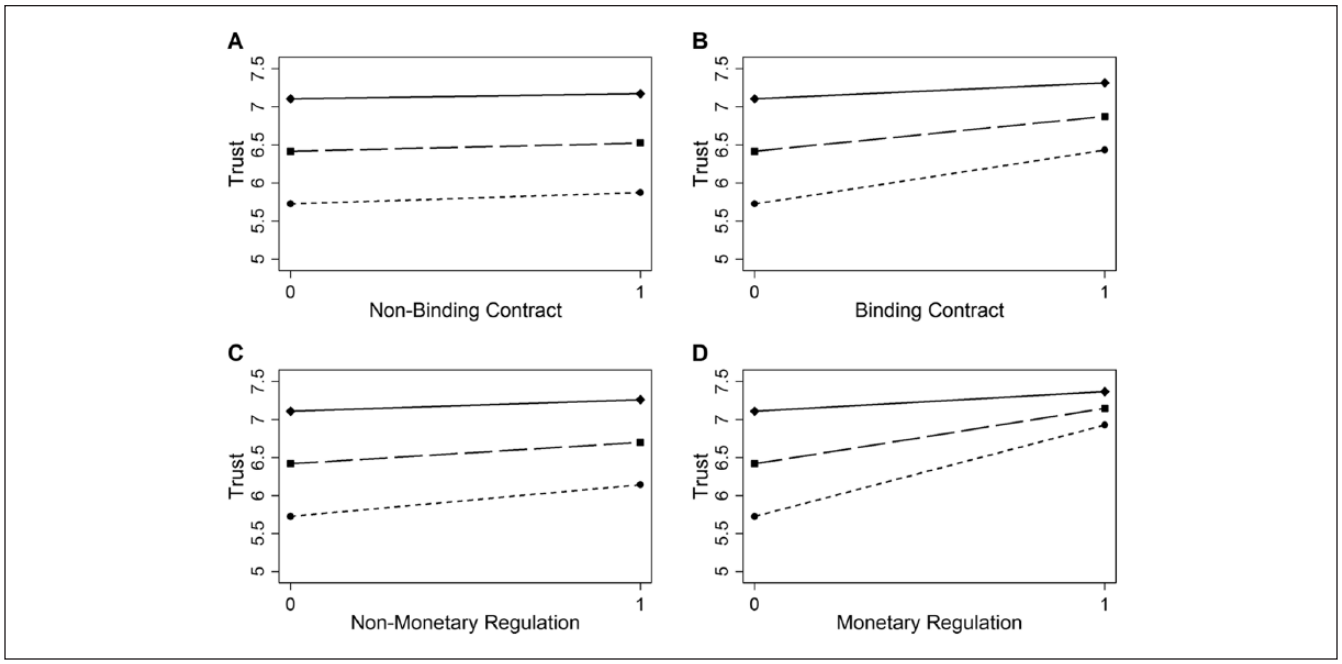

Figure 3. Relationship between social constraints and trust by causal attributions in the group project scenario (Table 1, model 4).
Note: For panels A to D, solid black slopes indicate dispositional attributions (10), dashed black slopes indicate both dispositional and situational attributions (5), and dotted black slopes indicate situational attributions (0).

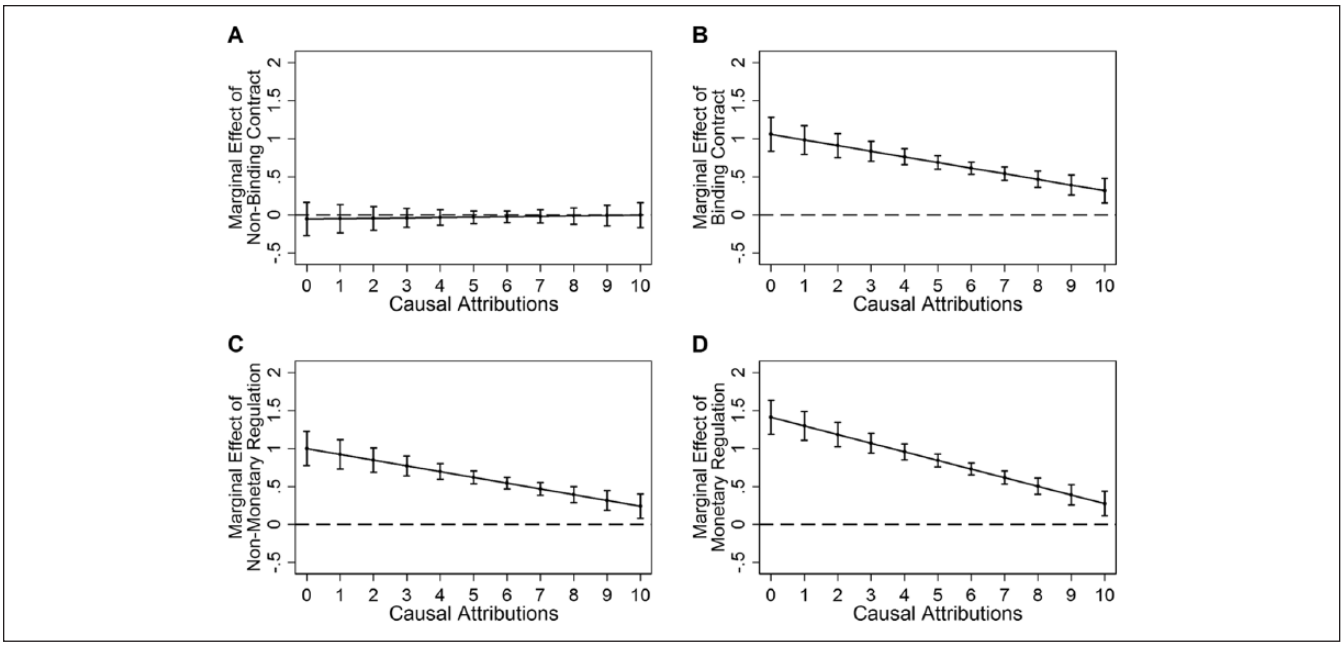

Figure 4. Marginal effects of social constraints by causal attributions in the car repair scenario (Table 1, model 2).

regulations on trust decreases as dispositional attributions increase. These effects were statistically significant at all values of the causal attributions variable, despite the statistical insignificance of the interaction term at the $p < .05$ level (see model 4).

Robustness Checks

I subjected the findings to a number of robustness checks. First, the substantive findings presented in Table 1 are robust to the exclusion of respondents who failed the screener questions, who partially completed an experiment, who

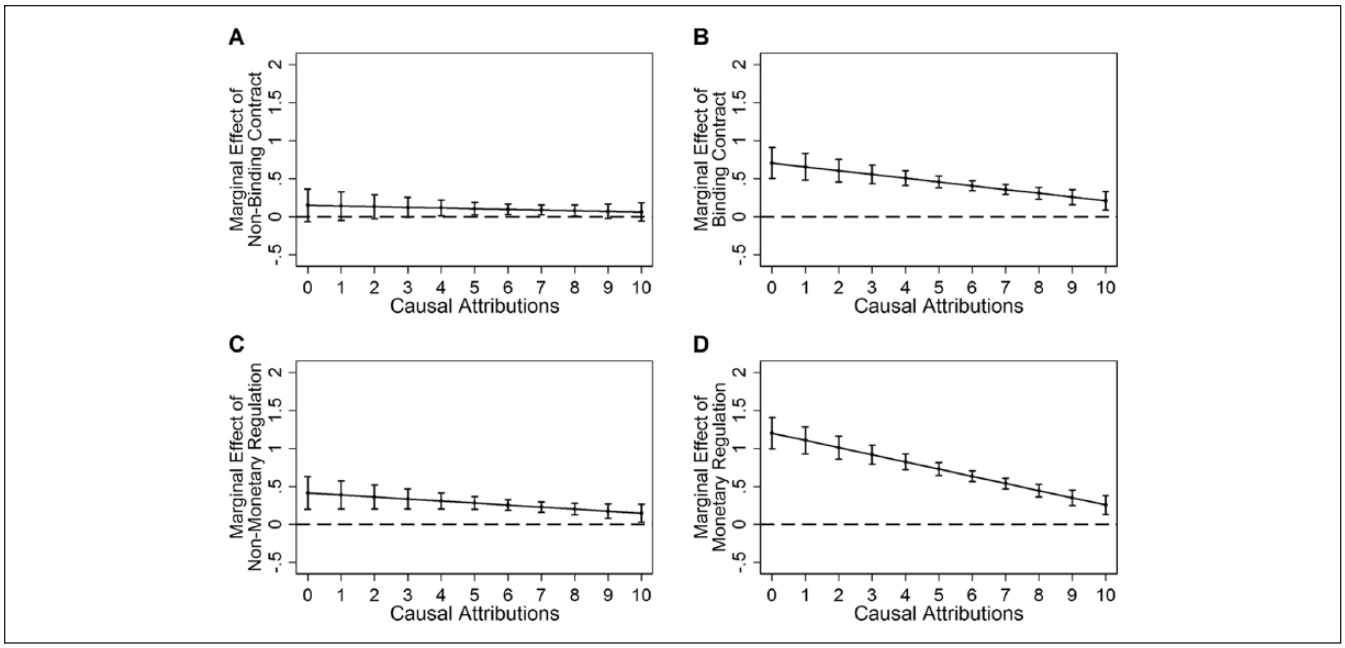

Figure 5. Marginal effects of social constraints by causal attributions in the group project scenario (Table 1, model 4).

participated in multiple experiments from the same IP address, or any combination of all three (see the subsection "Individual-level Covariates"). Second, the substantive findings presented in Table 1 are also robust to models in which X_{ij} and A_{ij} were not decomposed into within- and between-individual components (see equations 3 and 4). The results of these alternative model specifications can be found in the Supplemental Materials online.[6]

Discussion

Results from two large-*n* survey experiments yielded several key findings related to how social constraints influence trust. First, across both studies I found social constraints to increase trust. Trust significantly grew with the introduction of binding contracts and monetary regulations, and less so with nonbinding contracts and nonmonetary regulations. Second, as predicted, each type of social constraint, with the exception of nonbinding contracts, increased trust to the extent that individuals attributed another's perceived trustworthiness to the situation. But as individuals increasingly attributed another's perceived trustworthiness to disposition, the positive effect of social constraints on trust declined. Third, regardless of the presence or absence of social constraints (or the type of social constraint), reported levels of trust were always greater for individuals who drew dispositional versus situational attributions of another's perceived trustworthiness.

These results shed new light on the debate about whether trust blossoms (Farrell 2009; Greif 2006; Knight 2001; North 1990; Rothstein and Stolle 2008) or withers (Irwin et al. 2014; Malhotra and Murnighan 2002; Mulder et al. 2006; Simpson and Eriksson 2009) in the presence of social constraints. Overall, I find strong support for my integrated model. Social constraints promote trust but only under certain conditions (external attributions). And social constraints do not crowd-out trust per se, but instead produce null to weak positive effects under a different set of conditions (internal attributions). Attribution biases, then, are central to how social constraints influence trust. To my knowledge, this research is the first to document these particular dynamics, which is theoretically and empirically important given how causal attributions are treated as unmeasured and unobserved mediators in the crowding-out literature. Yet when a variable is assumed to mediate—but in reality moderates—one variable's effect on another, estimates of total effects are biased and inconsistent if said moderation is not statistically taken into account (Imai, Keels, and Tingley 2010). Future scholarship from the crowding-out perspective should synthesize the causal pathways proposed in my integrated model. Two possible routes exist. One route construes causal attributions as simultaneously mediating *and* moderating the effects of social constraints on trust. This solution, however, is not sensible or plausible (see Hayes 2013). The other route requires moderated mediation in which causal attributions moderate the indirect effects of a newly identified mechanism linking social constraints to trust. This latter solution is possible, but what this new mechanism might be I leave for future work.

[6]Also note that excluding vignettes with the uncooperative level does not alter the substantive findings presented in Table 1 (results are available upon request).

Although the pattern of findings is robust across scenarios, a challenge for my statistical model involves the post-treatment nature of causal attributions. The primary concern is that causal attributions, like trust, were neither directly manipulated nor randomly assigned to individuals. This produces a statistical problem in which it is impossible to determine whether causal attributions moderate the constraints-trust link or trust moderates the constraints-attributions link. Regarding this and other issues related to modeling interaction effects, Hargens (2009) suggested the following: (1) rely on theory to identify which variable moderates the other, (2) theoretically outline a mechanism(s) that generates moderation, and (3) at some point empirically measure said mechanism. In the present manuscript, I accomplish (1) and (2) but leave (3) for future research. Even though causal attributions are endogenous and their moderating effects possibly biased, the present research design provides a worthwhile and plausible test of my integrated model. This is akin to how observational data can deliver existence proofs and spur the development of research designs concerned with causal identification. The primary goal of my article was to follow a similar path: provide an existence proof as a first step into a larger research program.

Beyond fundamental statistical challenges, a number of avenues for future research exist. First, *lack of awareness* and *unrealistic expectations* were identified as possible sources of how and why some individuals underestimate the motivating power of social constraints. I did not, however, directly measure either of these inputs. Future work showing direct evidence of each would be a welcome addition to the literature. Second, given the hypothetical nature of my research design, a replication of the present findings using tangible stakes and concrete social constraints in a laboratory setting is an important next step. Third, the crowding-out effect observed in laboratory experiments (Irwin et al. 2014; Mulder et al. 2006; Simpson and Eriksson 2009) diverges from survey-based analyses of the state-trust relationship (Bjørnskov 2007; Delhey and Newton 2005; Herreros 2004; Robbins 2011, 2012a, 2012b). Although the present work adds a nuanced understanding of how political institutions might render or erode trust, future research should nonetheless investigate this cleavage in greater detail. Fourth, and finally, the logic of the crowding-out tradition has been applied to other related causes (e.g., reputations) and effects (e.g., generalized trust and trustworthiness) (Kuwabara 2015; Simpson and Eriksson 2009). Given the present findings, it would be advisable for future research to explore whether and to what extent causal attributions moderate these observed relationships.

To conclude, I find that social constraints increase trust. But the positive effect is conditional and does not always occur, thereby supporting my integrated model. Ultimately, those who attribute another's perceived trustworthiness to dispositional factors produce greater trust than those who do not, even in the presence of social constraints.

Author's Note

An earlier version of this article won the 2014 Outstanding Graduate Student Paper Award from the Altruism, Morality, and Social Solidarity section of the American Sociological Association.

Acknowledgments

I thank Maria Grigoryeva, Edgar Kiser, Ross Matsueda, Jerald Herting, Darryl Holman, Steven Pfaff, Lisa Keister, James Moody, and the anonymous reviewers for comments and suggestions; the Department of Sociology and the Graduate School at the University of Washington for funding support; and Richard Callahan for coding assistance. I also benefited from the opportunity to present parts of this work to the School of Information at the University of California, Berkeley.

Funding

This research was supported by a grant from the National Science Foundation (SES- 1303577), which bears no responsibility for the analysis and interpretations drawn here. Partial support for this research came from a Eunice Kennedy Shriver National Institute of Child Health and Human Development research infrastructure grant to the Center for Studies in Demography & Ecology at the University of Washington (R24 HD042828). Publication was made possible in part by support from the Berkeley Research Impact Initiative (BRII) sponsored by the UC Berkeley Library.

References

Auspurg, Katrin, and Thomas Hinz. 2015. *Factorial Survey Experiments*. Thousand Oaks, CA: Sage.

Berinsky, Adam J., Michele F. Margolis, and Michael W. Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-administered Surveys." *American Journal of Political Science* 58(3):739–53.

Bjørnskov, Chrisian. 2007. "Determinants of Generalized Trust: A Cross-country Comparison." *Public Choice* 130(1):1–21.

Chen, Xiao-Ping, Madan M. Pillutla, and Xin Yao. 2009. "Unintended Consequences of Cooperation Inducing and Maintaining Mechanisms in Public Goods Dilemmas: Sanctions and Moral Appeals." *Group Processes and Intergroup Relations* 12(2):241–55.

Cook, Karen S., Russell Hardin, and Margaret Levi. 2005. *Cooperation without Trust?* New York: Russell Sage.

Deci, Edward L. 1971. "Effects of Externally Mediated Rewards on Intrinsic Motivation." *Journal of Personality and Social Psychology* 18(1):105–15.

Deci, Edward L., Richard Koestner, and Richard M. Ryan. 1999. "A Meta-analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation." *Psychological Bulletin* 125(6):627–68.

Delhey, Jan, and Kenneth Newton. 2005. "Predicting Cross-national Levels of Social Trust: Global Pattern or Nordic Exceptionalism?" *European Sociological Review* 21(4):311–27.

Farrell, Henry. 2009. *The Political Economy of Trust: Institutions, Interests, and Inter-firm Cooperation in Italy and Germany*. New York: Cambridge University Press.

Gilbert, Daniel T., and Patrick S. Malone. 1995. "The Correspondence Bias." *Psychological Bulletin* 117(1):21–38.

Greif, Avner. 2006. *Institutions and the Path to the Modern Economy*. Cambridge, UK: Cambridge University Press.
Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22(1):1–30.
Hardin, Russell. 2002. *Trust & Trustworthiness*. New York: Russell Sage.
Hargens, Lowell. 2009. "Product-variable Models of Interaction Effects and Causal Mechanisms." *Social Science Research* 38(1):19–28.
Hayes, Andrew F. 2013. *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-based Approach*. New York: Guilford.
Herreros, Francisco. 2004. *The Problems of Forming Social Capital: Why Trust?* New York: Palgrave.
Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A General Approach to Causal Mediation Analysis." *Psychological Methods* 15(4):309–34.
Irwin, Kyle, Laetitia Mulder, and Brent Simpson. 2014. "The Detrimental Effects of Sanctions on Intra-group Trust: Comparing Punishments and Rewards." *Social Psychology Quarterly* 77(3):253–72.
Jasso, Guillermina. 2006. "Factorial Survey Methods for Studying Beliefs and Judgments." *Sociological Methods & Research* 34(3):334–423.
Jones, Edward E., and Victor A. Harris. 1967. "The Attribution of Attitudes." *Journal of Experimental Social Psychology* 3(1):1–24.
Knight, Jack. 2001. "Social Norms and the Rule of Law: Fostering Trust in a Socially Diverse Society." In *Trust & Society*, edited by Karen S. Cook. New York: Russell Sage.
Kuwabara, Ko. 2015. "Do Reputation Systems Undermine Trust? Divergent Effects of Enforcement Type on Generalized Trust and Trustworthiness." *American Journal of Sociology* 120(5):1390–1428.
Malhotra, Deepak, and J. Keith Murnighan. 2002. "The Effects of Contracts on Interpersonal Trust." *Administrative Science Quarterly* 47(3):534–59.
Mayer, Roger C., James H. Davis, and F. David Schoorman. 1995. "An Integrative Model of Organizational Trust." *Academy of Management Review* 20(3):709–34.
Morris, Michael W., and Kaiping Peng. 1994. "Culture and Cause: American and Chinese Attributions for Social and Physical Events." *Journal of Personality and Social Psychology* 67(6):949–71.
Mulder, Laetitia B., Eric Van Dijk, David De Cremer, and Henk A. M. Wilke. 2006. "Undermining Trust and Cooperation: The Paradox of Sanctioning Systems in Social Dilemmas." *Journal of Experimental Social Psychology* 42(2):147–62.
Mundlak, Yair. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica* 46(1):69–85.
Nannestad, Peter. 2008. "What Have We Learned about Generalized Trust, if Anything?" *Annual Review of Political Science* 11:413–36.
North, Douglas. 1990. *Institutions, Institutional Change, and Economic Performance*. New York: Cambridge University Press.
Putnam, Robert. 2000. *Bowling Alone: The Collapse and Revival of American Community*. New York: Free Press.
Robbins, Blaine. 2011. "Neither Government nor Community Alone: A Test of State-centered Models of Generalized Trust." *Rationality and Society* 23(3):304–46.
Robbins, Blaine. 2012a. "A Blessing and a Curse? Political Institutions in the Growth and Decay of Generalized Trust: A Cross-national Panel Analysis, 1980–2009." *PLoS ONE* 7(4):e35120.
Robbins, Blaine. 2012b. "Institutional Quality and Generalized Trust: A Nonrecursive Causal Model." *Social Indicators Research* 107(2):235–58.
Robbins, Blaine. 2016. "From the General to the Specific: How Social Trust Motivates Relational Trust." *Social Science Research* 55(1):16–30.
Robbins, Blaine. Forthcoming-a. "Probing the Links between Trustworthiness, Trust, and Emotion: Evidence from Four Survey Experiments." *Social Psychology Quarterly*.
Robbins, Blaine. Forthcoming-b. "What Is Trust? A Multidisciplinary Review, Critique, and Synthesis." *Sociology Compass*.
Ross, Lee. 1977. "The Intuitive Psychologist and His Shortcomings." Pp. 173–220 in L. Berkowitz (Ed.), *Advances in Experimental Social Psychology*, Vol. 10. San Diego, CA: Academic Press.
Rossi, Peter H., and Steven Knock. 1982. *Measuring Social Judgments: The Factorial Survey Approach*. Beverly Hills, CA: Sage.
Rothstein, Bo. 2000. "Trust, Social Dilemmas and Collective Memories." *Journal of Theoretical Politics* 12(4):477–501.
Rothstein, Bo, and Dietlind Stolle. 2008. "How Political Institutions Create and Destroy Social Capital: An Institutional Theory of Generalized Trust." *Comparative Politics* 40(4):441–59.
Schunck, Reinhard. 2013. "Within and Between Estimates in Random-effects Models: Advantages and Drawbacks of Correlated Random Effects and Hybrid Models." *Stata Journal* 13(1):65–76.
Simpson, Brent, and Kimmo Eriksson. 2009. "The Dynamics of Contracts and Generalized Trustworthiness." *Rationality and Society* 21(1):59–80.
Sztompka, Piotr. 1999. *Trust: A Sociological Theory*. Cambridge, UK: Cambridge University Press.
Titmuss, Richard. 1970. *The Gift Relationship: From Human Blood to Social Policy*. New York: New Press.
Uslaner, Eric. 2002. *The Moral Foundations of Trust*. Cambridge, UK: Cambridge University Press.
Wilson, Rick K., and Catherine C. Eckel. 2011. "Trust and Social Exchange." In *Cambridge Handbook of Experimental Political Science*, edited by J. N. Druckman, D. P. Green, J. H. Kuklinski, and A. Lupia. Cambridge, UK: Cambridge University Press.
Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.
Yamagishi, Toshio, and Midori Yamagishi. 1994. "Trust and Commitment in the United States and Japan." *Motivation and Emotion* 18(2):129–66.

Author Biography

Blaine G. Robbins is an assistant professor of social research and public policy at New York University Abu Dhabi. His research focuses on trust and trustworthiness, collective action, social norms, and network formation. His work has appeared in the *Journal of Cross-cultural Psychology, Rationality and Society, Social Psychology Quarterly, Social Science Research*, and other peer-reviewed journals. Current and ongoing projects include the analysis of betrayal, the causes of tax evasion, and the origins of social predation.