Additional Materials

Social Psychology Quarterly

Probing the Links between Trustworthiness, Trust, and Emotion:
Evidence from Four Survey Experiments

Date Created: 6-9-2016
Date Modified: 6-9-2016

Table of Contents

Table A1. Sample Descriptive Statistics, Studies 1a and 1b

		Sample Descriptives			
Name	Definition	Study 1a	N	Study 1b	N
Age	Age in years (study 1a: 18 to 81; study 1b: 18 to 76)	32.61 (11.51)	1378	32.10 (10.89)	1403
Race					
Non-Hispanic white	Non-Hispanic white respondent	75%	1376	71%	1406
Non-white Hispanic	Non-white Hispanic respondent	1%	1376	2%	1406
Non-white black	Non-white black respondent	5%	1376	7%	1406
Non-white Asian	Non-white Asian respondent	5%	1376	5%	1406
Other race	Other race (includes DK/PNS)	14%	1376	15%	1406
Gender					
Female	Female respondent	48%	1369	56%	1388
Male	Male respondent	52%	1369	44%	1388
Other	Other gender respondent (includes PNS)	< 1%	1369	< 1%	1388
Education					
Less than high school	No degree received	1%	1377	1%	1405
High school	Highest degree received is HS diploma or GED	38%	1377	33%	1405
Some college	Highest degree received is AA or equivalent	15%	1377	14%	1405
College or more	Highest degree received is at least a bachelor's degree	46%	1377	51%	1405
DK/PNS	Highest degree received is DK or PNS	< 1%	1377	1%	1405
Income					
Less than $24,999	Household income is less than or equal to $24,999 a year	24%	1373	25%	1402
$25,000-$49,999	Household income is between $25,000 and $49,999 a year	32%	1373	27%	1402
$50,000-$74,999	Household income is between $50,000 and $74,999 a year	20%	1373	18%	1402
Greater than $75,000	Household income is greater than $75,000 a year	20%	1373	22%	1402
DK/PNS	Household income is DK or PNS	4%	1373	6%	1402
Marital status					
Never married	Respondent has never married	59%	1373	54%	1396
Married	Respondent is married	31%	1373	34%	1396
Other	Respondent is either separated, divorced, or widowed	9%	1373	11%	1396

PNS	Respondent is PNS	1%	1373	1%	1396
Religion					
Catholic	Respondent identifies as Catholic	13%	1371	13%	1403
Mainline Protestant	Respondent identifies as Baptist, Lutheran, Methodist, or Presbyterian	13%	1371	13%	1403
Christian, non-denom.	Respondent identifies as Christian, non-denominational	8%	1371	10%	1403
No religion	Respondent identifies as 'no religion'	33%	1371	25%	1403
No affiliation	Respondent does not identify with a particular religious group	15%	1371	17%	1403
Other affiliation	Respondent identifies with some other religious group	16%	1371	20%	1403
DK	Religious group affiliation is DK	2%	1371	2%	1403
Evangelical					
Yes	Respondent identifies as an Evangelical Christian	10%	1375	11%	1405
No	Respondent does not identify as an Evangelical Christian	83%	1375	79%	1405
DK/PNS	Respondent is DK or PNS for Evangelical Christian	7%	1375	10%	1405
Political ideology	Political ideology scale (1 to 7; extremely liberal to extremely conservative)	3.40	1302	3.43	1277
		(1.66)		(1.64)	

Note: Standard deviations in parentheses for continuous variables. DK = don't know, PNS = prefer not to say.

Table A2. Sample Descriptive Statistics, Studies 2a and 2b

		Sample Descriptives			
Name	Definition	Study 2a	N	Study 2b	N
Age	Age in years (study 2a: 18 to 63; study 2b: 18 to 60)	20.78	924	20.79	883
		(4.41)		(4.27)	
Race					
Non-Hispanic white	Non-Hispanic white respondent	50%	927	46%	887
Non-white Hispanic	Non-white Hispanic respondent	3%	927	3%	887
Non-white black	Non-white black respondent	1%	927	1%	887
Non-white Asian	Non-white Asian respondent	25%	927	28%	887
Other race	Other race (includes DK/PNS)	21%	927	22%	887
Gender					
Female	Female respondent	61%	919	64%	875
Male	Male respondent	39%	919	36%	875
Other	Other gender respondent (includes PNS)	< 1%	919	< 1%	875
Education					
Less than high school	No degree received	2%	923	2%	884
High school	Highest degree received is HS diploma or GED	71%	923	73%	884
Some college	Highest degree received is AA or equivalent	21%	923	20%	884
College or more	Highest degree received is at least a bachelor's degree	5%	923	4%	884
DK/PNS	Highest degree received is DK or PNS	1%	923	1%	884
Income					
Less than $24,999	Household income is less than or equal to $24,999 a year	27%	918	29%	879
$25,000-$49,999	Household income is between $25,000 and $49,999 a year	10%	918	11%	879
$50,000-$74,999	Household income is between $50,000 and $74,999 a year	8%	918	8%	879
Greater than $75,000	Household income is greater than $75,000 a year	29%	918	26%	879
DK/PNS	Household income is DK or PNS	26%	918	26%	879
Marital status					
Never married	Respondent has never married	95%	925	96%	880
Married	Respondent is married	3%	925	2%	880
Other	Respondent is either separated, divorced, or widowed	1%	925	1%	880

PNS	Respondent is PNS	1%	925	1%	880
Religion					
Catholic	Respondent identifies as Catholic	11%	922	12%	876
Mainline Protestant	Respondent identifies as Baptist, Lutheran, Methodist, or Presbyterian	7%	922	6%	876
Christian, non-denom.	Respondent identifies as Christian, non-denominational	11%	922	11%	876
No religion	Respondent identifies as 'no religion'	28%	922	31%	876
No affiliation	Respondent does not identify with a particular religious group	21%	922	18%	876
Other affiliation	Respondent identifies with some other religious group	18%	922	17%	876
DK	Religious group affiliation is DK	4%	922	5%	876
Evangelical					
Yes	Respondent identifies as an Evangelical Christian	11%	925	7%	876
No	Respondent does not identify as an Evangelical Christian	76%	925	80%	876
DK/PNS	Respondent is DK or PNS Evangelical Christian	13%	925	13%	876
Political ideology	Political ideology scale (1 to 7; extremely liberal to extremely conservative)	3.15	778	3.14	729
		(1.45)		(1.34)	

Note: Standard deviations in parentheses for continuous variables. DK = don't know, PNS = prefer not to say.

AMAZON.COM'S MECHANICAL TURK: STUDIES 1AB

Online experiments in which subjects are recruited from "crowdsource" labor sites have become more common in the social sciences. On these sites, workers are compensated for (usually) small tasks, such as finding Waldo in a "*Where's Waldo?*" image or participating in a social science experiment. The largest and most common of these crowdsource labor sites is Amazon.com's Mechanical Turk (MTurk) where requesters post jobs—or Human Intelligence Tasks (HITs)—and workers choose which jobs (or HITs) to do for pay.

The goal of this section is not to review the "ins-and-outs" of how to use MTurk for social science experiments; the goal is to provide the reader with more information about how data were collected on MTurk, what steps I took to reduce bias, and the literature detailing the validity and reliability of social sciences experiments conducted on MTurk. For those who are interested in more information about MTurk and how to conduct experiments on MTurk, I refer the reader to Mason and Suri (2012) and Paolacci et al. (2010); for recent empirical pieces using MTurk in sociology, please see Abascal (2015), Harrell and Simpson (2016), Kuwabara (2015), Kuwabara and Sheldon (2012), and Simpson et al. (2013); and see Shank (2016) for a review of this literature.

Data Collection on MTurk

For studies 1a and 1b, participants were contacted over Amazon.com's Mechanical Turk via a HIT advertising $2 payment for participating in a 20-minute web-based experiment.[1] Studies 1a and 1b were published on Amazon.com's Mechanical Turk Requester site on October 2nd and October 9th of 2013, respectively. The HIT was called "Complete a [Institution Blinded] survey

[1] By MTurk standards, $2 is relatively high for a 20-minute HIT. $2 compensation was used to increase study completion time.

about perceptions of trustworthiness" and described the task as research on trust and cooperation. While the description was vague, I provided keywords to describe the HIT as *survey*, *research*, *social psychology*, *demographics*, and *opinion*. When workers opened the HIT, I instructed them to click on a survey URL link, which directed them to my Limesurvey platform.

To be eligible, MTurk workers must have been legal adults residing in the U.S. with approval rates 90% or above on previous MTurk tasks—that is, previous Requesters accepted 90% or more of the HITs submitted by a worker.[2] In total, 1,388 Amazon.com Mechanical Turk (MTurk) workers participated in study 1a from 1,340 unique IP addresses (IPs), while 1,354 of these workers completed study 1a from 1,311 unique IPs. Likewise, 1,419 MTurk workers participated in study 1b from 1,388 unique IPs, while 1,369 of these workers completed study 1b from 1,342 unique IPs (for descriptive statistics, see Table A1). In both studies, workers were paid $2 for completing the study. Each study consisted of a unique Mechanical Turk sample with no redundant workers across samples. Overall sample size was determined by a power analysis, and my data collection stopping rule for each study consisted of reaching a target sample size of 1,350 respondents who completed *and* submitted a HIT to Amazon.com.[3] For

[2] To maximize data quality and minimize attrition, Berinsky et al. (2012) and Mason and Suri (2012) suggest restricting eligibility to workers with approval rates of at least 90% or above on previous MTurk tasks.

[3] The data collection stopping rule of 1,350 completed and submitted HITs was exceeded by 4 in study 1a and 19 in study 1b. There are two reasons for this discrepancy. First, upon completing the web-based survey experiment, MTurk workers were assigned a unique survey code to submit as a completed HIT to Amazon.com. In some cases, workers completed the experiment but did not submit their HIT. Second, some workers completed the experiment but were barred from submitting their HIT by Amazon.com since the target goal of 1,350 submitted HITs was achieved. In all of these instances, workers who

both studies, the target sample size of 1,350 respondents was reached in two days. The median time to complete studies 1a and 1b was 18.12 and 18.87 minutes, respectively.

Bias Reduction on MTurk

I now provide an account of why MTurk is appropriate for the current study and the steps I took to reduce bias. First, although MTurk samples frequently yield distributions of demographic characteristics that parallel those found with representative population-based random samples (Berinsky et al. 2012; Weinberg et al. 2014), data drawn from MTurk are based on non-probabilistic, convenience samples. This comes with the usual caveats: convenience samples are usually unrepresentative and, as a result, undermine external validity and population-based inferences. But since I am interested in theoretical, model-based inferences instead of population-based inferences (much like other lab experimentalists who test theory by drawing on undergraduate student populations), this factor is less of an issue for internal validity. Finally, the demographic characteristics found in Table A1 parallel those found in prior MTurk studies (see Berinsky et al. 2012; Shapiro et al. 2013; Weinberg et al. 2014).

Second, MTurk workers are known to share information about HITs and how to respond to them (Chandler et al. 2014). Such behavior violates the classic "stable unit treatment value" assumption (or SUTVA), which can bias parameter estimates if violated. To investigate whether SUTVA was violated, I searched known MTurk forums and discussion threads to see what was discussed about the current study's HITs and when. I discovered nothing other than discussions about my HITs paying well given the time it took to complete the HITs. I also tried to minimize

---

completed the experiment but failed to or were unable to submit their HIT were compensated $2 upon contacting the researcher. Importantly, the results presented in studies 1a and 1b do not substantively change when excluding those workers who completed the experiment but failed to submit their HIT.

the violation of SUTVA by paying workers well enough to finish data collection as soon as possible (i.e., in two days), since HITs posted for longer periods of time have a greater probability of being discussed among potential respondents. Violation of SUTVA was also dealt with by including dummy variables for IP address in the level-2 model (see *SPQ*'s Supplemental Materials online).

Third, research has found that MTurk harbors a class of workers who gravitate toward taking surveys and participating in experiments (Chandler et al. 2014). Furthermore, this research shows that many workers are familiar with "off-the-shelf" or "canned" behavioral experiments such as prisoner's dilemma games and investment (or trust) games. Researchers interested in naïve subjects should avoid using such common experimental procedures on MTurk. I created and implemented two novel survey experiments to circumvent this issue.

Fourth, research suggests that screener questions should be used to assess the attention of MTurk workers (Berinsky et al. 2014). If workers are inattentive, then these workers will receive but not cognitively process the treatment(s). This yields outcomes similar to failure-to-treat or one-sided noncompliance (see Gerber and Green 2012, ch. 5), which can ultimately bias estimates of treatment effects. To address this issue, respondents read a coversheet detailing the hypothetical scenario and the task at hand (the design also included a CAPTCHA of "What is 3 + 5?" during the consent process that was used to exclude non-human participants). Just after the coversheet but prior to the 10 vignettes, respondents were quizzed about material embedded in the coversheet. This quiz consisted of two T/F questions. Results suggest that MTurk workers read and understood material found in the coversheet (study 1a: screener 1 = 96% correct, screener 2 = 98% correct; study 1b: screener 1 = 95% correct, screener 2 = 94% correct), which

implies that workers were attentive enough to comprehend the 10 dimensions across the 10 vignettes.

Fifth, and finally, I also address other issues, such as non-independence of errors, which can be found in *SPQ*'s Supplemental Materials online.

Validity and Reliability of MTurk Samples

In this sub-section, I review what prior research reveals about experiments and surveys administered on MTurk and how findings from this research parallel studies using representative population-based random samples and undergraduate student samples. First, with respect to the distribution of demographic characteristics, MTurk samples are very similar to representative population-based random samples. Berinsky et al. (2012) found that MTurk workers were slightly younger and more liberal than population-based random samples administered either via the internet or face-to-face. Women also tended to constitute a greater share of the MTurk population (roughly 60%), and median incomes for MTurk workers were about $10,000 less than population-based random samples (both internet and face-to-face). Other studies have found similar results and distributions of demographic characteristics on MTurk (Shapiro et al. 2013; Weinberg et al. 2014). The present study's data collection efforts also yielded similar distributions of demographic characteristics (see Table A1).

Second, recent research has replicated numerous experiments using MTurk samples. Berinsky et al. (2012) replicated (1) a classic study of the effect of question wording on survey responses (Rasinski 1989), (2) the canonical loss aversion experiment (Tversky and Kahneman 1981), and (3) a political science experiment on the effects of risk preferences on susceptibility to framing (Kam and Simas 2010). Likewise, Horton et al. (2011) replicated (a) the existence and extent of other-regarding preferences found in prior work (see Camerer 2003), (b) findings

from religious priming studies (Shariff and Norenzayan 2007), and (c) a classic framing experiment (Tversky and Kahneman 1981). In a third study, Suri and Watts (2011) replicated prior findings from repeated public goods games run in physical laboratories by administering 113 web-based experiments on MTurk over a 6 month period.

More recently, Weinberg et al. (2014) compared results of three survey experiments conducted on GfK's population-based online survey platform and MTurk (see also Mullinix et al. 2015). They showed that the results of their experiments were very similar and that indicators of data quality were slightly better among MTurk workers. Finally, Shapiro et al. (2013) have shown that "Consistent with earlier research on the psychometric properties of personality scales on MTurk (Buhrmester et al. 2011), mental health measures were found, overall, to demonstrate satisfactory internal reliability and test-retest reliability. Extending this work, we demonstrate the criterion validity of these measures on the MTurk population by replicating associations between psychopathology and established demographic predictors (e.g., unemployment) (p. 5)."

In short, MTurk can be used to rapidly obtain inexpensive high-quality data that parallels results of experiments and surveys obtained from representative population-based random samples as well as unrepresentative convenience-based samples (e.g., undergraduate student populations). In other words, the internal *and* external validity of experiments using MTurk samples is strong, with subjects recruited in this manner often more representative of the U.S. population than common undergraduate student samples but less representative than national probability samples (e.g., General Social Survey).

DESIGN-BASED ASSUMPTIONS AND ROBUSTNESS CHECKS: STUDIES 1AB

Design-Based Assumptions

First, under two weak assumptions, the assessment of ten vignettes per respondent allows me to estimate unit causal effects in a counterfactual framework and not merely treatment effects on the treated. The two assumptions state that potential outcomes are unaffected by the anticipation of treatments administered in the future *and* that potential outcomes in one period are unaffected by treatments administered in prior periods (see Holland 1986; Hainmueller et al. 2014). To test these assumptions, I estimated (a) separate regressions for each vignette subsample (yielding 10 models total for each of the i vignettes where $i = 1,\ldots,10$), (b) two-level models in which each vignette dimension, X_{ij}, interacted with V_i, the vignette dummy variables, and (c) two-level models with vignette dimensions and lagged t - 1 vignette dimensions. Consistent with these two assumptions, I found that (a) the parameter estimates for each vignette subsample were indeed similar in studies 1a and 1b, (b) a χ^2 test for the joint significance of the interaction terms could not reject the null that the vignette dimensions are identical across the 10 vignettes (study 1a: χ^2 (198) = 230.20, $p > .05$; study 1b: χ^2 (198) = 199.26, $p > .05$), and (c) a χ^2 test for the joint significance of the lagged t - 1 vignette dimensions failed to reject the null hypothesis in study 1b (χ^2 (22) = 30.80, $p > .05$) but not study 1a (χ^2 (22) = 55.25, $p < .001$).[4]

Second, while uncommon for laboratory experiments, it is common for—and even considered a strength of—survey experiments to consist of a large number of dimensions where respondents assess multiple vignettes (Auspurg and Hinz 2015; Hainmueller et al. 2015;

[4] Although the test of joint significance for the lagged t - 1 vignette dimensions was statistically significant in study 1a, familywise tests were statistically insignificant for study 1b and studies 2ab. I thus interpret the results of the lagged t - 1 familywise test for study 1a as statistical noise.

Hainmueller et al. 2014; Jasso 2006). With such designs, a key assumption of no order effects must hold and researchers should avoid other design-based threats to internal validity—such as fatigue effects and learning effects—to ensure unbiased estimates of treatment effects (see Hainmueller et al. 2014). Recent methodological research suggests that the current survey experiments—ten dimensions per vignette with ten vignettes assessed per person with simple evaluation tasks—fall within design bounds that minimize cognitive overload, learning effects, and order effects (Auspurg and Hinz 2015; Auspurg and Jäckle 2015; Sauer et al. 2011). Although order effects have been observed in survey experiments using more than one evaluation task (i.e., outcome variable), this bias is unique to complex evaluation tasks (i.e., open-ended responses) and is relatively minor and marginally significant (Auspurg and Jäckle 2015). Given this issue, it is advisable to minimize the number of evaluation tasks in a survey experiment where multiple vignettes are assessed with a large number of dimensions presented in a fixed order. As a result, I have chosen to assume zero measurement error for all dependent variables in favor of minimizing cognitive overload, learning effects, and order effects. To test for possible fatigue effects and other issues related to repeated measures (e.g., unobserved effects that might influence all cases to the same degree for a specified *i*th vignette), I regressed trust on V_i , the vignette dummy variables. A χ^2 test for the joint significance of the vignette dummy variables rejected the null that the vignette dummy variables are identical across the 10 vignettes (study 1a: $\chi^2(9) = 148.38$, $p < .05$; study 1b: $\chi^2(9) = 186.18$, $p < .05$). As a result, I included a vector of vignette dummy variables in the level-1 model (see eq. 1 in *SPQ*'s Supplemental Materials online).

Third, I checked whether randomization produced experimental groups that were well balanced in studies 1a and 1b. I assessed this assumption with multivariate balance checks by

regressing each vignette dimension—either logistic or multinomial logistic regression depending on the number of levels per dimension—on the individual-level covariates found in Table A1. I found that the individual-level covariates were jointly insignificant (p-values > .05) for each vignette dimension in studies 1a and 1b, indicating that the vignette dimensions were jointly balanced.

Fourth, with respect to other design-based considerations, none of the dimension combinations produced atypical or unrealistic vignettes (Faia 1980), which heightens concerns about external validity when present since unrealistic combinations deem counterfactuals meaningless. Since none of the dimensions produced unrealistic combinations, this issue is of little concern. Because vignette dimensions were randomly assigned with replacement it is possible for respondents to assess the same vignette by pure random chance. For studies 1a and 1b, none of the respondents assessed two of the same vignettes.

Fifth, since a simple random design with replacement was used (see Dülmer 2015; Rossi and Anderson 1982), the factorial object universe of 51,840 unique vignettes is greater than the overall sample size across all four experiments. This sort of design implies that main effects and higher order interaction effects are not necessarily orthogonal, which produces *aliasing* (Alexander and Becker 1978) or the confounding of main effects with higher order interaction effects (with perfect aliasing a higher order interaction effect cannot be separated statistically from a main effect). Since my primary theoretical interest is in main effects and lower order interaction effects, I assume that aliased higher order interaction effects are negligible relative to the affected main effects and lower order interaction effects (Gunst and Mason 1991; Dülmer 2015). This is a weak assumption given the consistent main effects and lower order interaction effects observed across all four experiments (where a unique vignette set is randomly produced

for each experiment). Finally, with this assumption, I have sufficient statistical power to test for main effects and lower order interaction effects.

Robustness Checks

As shown in Table A3, the substantive findings presented in Table 1 of the main document are robust to (a) specifications in which X_{ij} are not decomposed into within- and between-individual components (see models 1 through 4), and (b) the exclusion of respondents who failed the screener questions, who partially completed an experiment, who participated in multiple experiments from the same IP address, or any combination of all three (see models 3 through 6). Finally, excluding vignettes with the uncooperative level does not alter the substantive findings presented in Table 1 (the results of this alternative model specification are available upon request).

Table A3. Model Robustness Checks, Studies 1a and 1b

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
	Car Repair	Group Project	Car Repair	Group Project	Car Repair	Group Project
No prior interaction (ref.)						
Uncooperative	-2.73*** (.08)	-2.15*** (.07)	-2.79*** (.09)	-2.16*** (.07)	-2.82*** (.09)	-2.17*** (.07)
Prior interaction	1.13*** (.05)	1.20*** (.05)	1.16*** (.06)	1.25*** (.05)	1.15*** (.06)	1.24*** (.05)
Encapsulated interests	1.36*** (.05)	1.32*** (.05)	1.40*** (.06)	1.40*** (.05)	1.39*** (.06)	1.40*** (.05)
Goodwill	2.16*** (.05)	1.66*** (.05)	2.19*** (.06)	1.75*** (.05)	2.18*** (.06)	1.76*** (.05)
Virtuous dispositions	2.07*** (.06)	1.70*** (.05)	2.18*** (.06)	1.77*** (.05)	2.17*** (.06)	1.77*** (.05)
No contract (ref.)						
Non-binding contract	-0.04 (.04)	0.10** (.03)	-0.03 (.04)	0.10** (.04)	-0.03 (.04)	0.10** (.04)
Binding contract	0.59*** (.04)	0.36*** (.03)	0.65*** (.04)	0.37*** (.03)	0.67*** (.05)	0.38*** (.03)
No regulation (ref.)						
Non-monetary regulation	0.55*** (.04)	0.23*** (.03)	0.56*** (.04)	0.23*** (.04)	0.58*** (.04)	0.23*** (.04)
Monetary regulation	0.73*** (.04)	0.55*** (.03)	0.75*** (.04)	0.54*** (.04)	0.77*** (.04)	0.55*** (.04)
Constant	4.77*** (.35)	5.28*** (.28)	4.71*** (.29)	5.94*** (.24)	5.69*** (.49)	5.86*** (.51)
var(u_{0j})	0.74*** (.06)	0.73*** (.05)	0.83*** (.07)	0.76*** (.05)	0.78*** (.07)	0.73*** (.05)
var(e_{ij})	2.95*** (.07)	2.17*** (.06)	2.82*** (.07)	2.03*** (.05)	2.83*** (.07)	2.03*** (.06)
Other vignette dimensions	Yes	Yes	Yes	Yes	Yes	Yes
Vignette dummies	Yes	Yes	Yes	Yes	Yes	Yes
Individual-specific mean dim.	No	No	No	No	Yes	Yes
Individual-level covariates	Yes	Yes	Yes	Yes	Yes	Yes
Exclusions[1]	No	No	Yes	Yes	Yes	Yes
Observations	13733	14019	11349	11266	11319	11236
Individuals	1383	1414	1140	1131	1137	1128

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.10$ (two-tailed)

Note: unstandardized slopes (robust standard errors in parentheses).

[1] Excludes failed screeners, partial completes, and shared IP addresses.

PUBLIC UNIVERSITY UNDERGRADUATE STUDENTS: STUDIES 2AB

The goal of this section is to provide the reader with more information about how data for studies 2a and 2b were collected and what steps I took to reduce bias.

Data Collection

For studies 2a and 2b, 10,000 undergraduate students at a large public university were contacted over e-mail advertising a $50 lottery payment for participating in a 20-minute web-based experiment. Of these 10,000 students, 5,000 were randomly selected for studies 2a and 2b, respectively. Data collection for each study started on October 29th of 2013. The recruitment e-mail was called "Complete a [Department Blinded] survey for a chance to win $50" and described the study as research on trust and cooperation. The body of the e-mail outlined the study (e.g., how people come to believe that others are trustworthy), the expected study length (20 minutes), the students' rights as subjects (e.g., free not to answer any question they do not wish to answer), and the form of compensation (a chance to win one of six $50 bills). At the bottom of the recruitment e-mail, I instructed students how to participate or opt-out of the study by clicking on a survey URL link, which directed them to my Limesurvey platform. Students were also provided with a study-specific e-mail for questions, inquires, or additional information

To be eligible, students must have been legal adults currently enrolled at the [University blinded]. In total, 995 undergraduate students participated in study 2a from 987 unique survey tokens and 964 unique IP addresses (IPs), while 895 of these students completed study 2a from 895 unique survey tokens and 877 unique IPs. Likewise, 956 undergraduate students participated in study 2b from 946 unique survey tokens and 926 unique IPs, while 851 of these students completed study 2b from 851 unique survey tokens and 836 unique IPs (for descriptive statistics, see Table A2). With respect to response rates, 13 students did not meet study 2a's

eligibility requirements (e.g., reported not being legal adults), which produced an 18% response rate (895 completed/4,987 eligible); while 16 students did not meet study 2b's eligibility requirements, which produced a 17% response rate (851 completed/4,984 eligible). In both studies, students were entered into a lottery for the chance to win one of six $50 bills. Each study consisted of a unique student sample with no redundant students across samples. Overall sample size was determined by a power analysis, and my data collection stopping rule consisted of one full academic quarter of recruitment with one recruitment e-mail sent a week. The median time to complete studies 2a and 2b was 19.57 and 19.65 minutes, respectively.

Bias Reduction

First, the possible violation of SUTVA was dealt with by including dummy variables for IP address in the level-2 model (see *SPQ*'s Supplemental Materials online). Second, following studies 1a and 1b, just after the coversheet but prior to the 10 vignettes, respondents were quizzed about material embedded in the coversheet. Again, this quiz consisted of two T/F questions. Results suggest that the undergraduate students read and understood material found in the coversheet (study 2a: screener 1 = 90% correct, screener 2 = 96% correct; study 2b: screener 1 = 93% correct, screener 2 = 93% correct), which implies that workers were attentive enough to comprehend the 10 dimensions across the 10 vignettes (but less so than the MTurk workers found in studies 1a and 1b). Third, and finally, I address other issues, such as non-independence of errors, which can be found in *SPQ*'s Supplemental Materials Online.

DESIGN-BASED ASSUMPTIONS AND ROBUSTNESS CHECKS: STUDIES 2AB

Design-Based Assumptions

First, to test the assumptions that potential outcomes are unaffected by the anticipation of treatments administered in the future *and* that potential outcomes in one period are unaffected by treatments administered in prior periods, I estimated (a) separate regressions for each vignette subsample (yielding 10 models total for each of the *i* vignettes where *i* = 1,…,10), (b) two-level models in which each vignette dimension, X_{ij}, interacted with V_i, the vignette dummy variables, and (c) two-level models with vignette dimensions and lagged t - 1 vignette dimensions. Consistent with these two assumptions, I found that (a) the parameter estimates for each vignette subsample were indeed similar in studies 2a and 2b, (b) a χ^2 test for the joint significance of the interaction terms could not reject the null that the vignette dimensions are identical across the 10 vignettes in study 2a (χ^2 (198) = 204.30, $p > .05$) but not study 2b (χ^2 (198) = 243.57, $p = .02$), and (c) a χ^2 test for the joint significance of the lagged t - 1 vignette dimensions failed to reject the null hypothesis in both studies (study 2a: χ^2 (22) = 23.24, $p > .05$; study 2b: χ^2 (22) = 20.31, $p > .05$) . Regarding study 2b in (b), a post-hoc analysis revealed that the statistically significant omnibus test was driven by an interaction between the Competence dimension and V_i. Given that similar effects were not observed in the other studies and that the AIC and BIC were smaller in a model without the X_{ij} and V_i interactions, it is safe to conclude that these two assumptions have not been violated in study 2a or study 2b.

Second, with respect to cognitive overload, learning effects, and order effects, studies 2ab use the exact same design as studies 1ab but with slightly different dependent variables. As a result, the survey experiments used in studies 2ab fall within design bounds that minimize cognitive overload, learning effects, and order effects. To test for possible fatigue effects and

other issues related to repeated measures, I regressed trust on V_i , the vignette dummy variables. A χ^2 test for the joint significance of the vignette dummy variables rejected the null that the vignette dummy variables are identical across the 10 vignettes (study 2a: $\chi^2(9) = 91.06$, $p < .05$; study 2b: $\chi^2(9) = 45.05$, $p < .05$). As a result, I included a vector of vignette dummy variables in the level-1 model (see eq. 3 in *SPQ*'s Supplemental Materials online).

Third, regarding whether randomization produced experimental groups that were well balanced in studies 2a and 2b, I regressed each vignette dimension—either logistic or multinomial logistic regression depending on the number of levels per dimension—on the individual-level covariates found in Table A2. I found that the individual-level covariates were jointly insignificant (p-values > .05) for each vignette dimension in studies 2a and 2b, indicating that the vignette dimensions were jointly balanced.

Fourth, with respect to other design-based considerations, none of the respondents assessed two of the same vignettes in study 2a, while only one respondent in study 2b assessed two of the same vignettes. Excluding this respondent does not alter the results presented in the main document.

Robustness Checks

As shown in Table A4, the substantive findings presented in Table 2 of the main document are robust to (a) specifications in which X_{ij} are not decomposed into within- and between-individual components (see models 1 through 4), and (b) the exclusion of respondents who failed the screener questions, who partially completed an experiment, who participated in multiple experiments from the same IP address, or any combination of all three (see models 3 through 6). The (a) and (b) robustness checks outlined above also apply to the mediation analysis found in Tables 3 and 4 (results available upon request). Finally, excluding vignettes with the

uncooperative level does not alter the substantive findings presented in Tables 2 through 4 (the results of these alternative model specifications are available upon request).

Table A4. Model Robustness Checks, Studies 2a and 2b

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
	Car Repair	Group Project	Car Repair	Group Project	Car Repair	Group Project
No prior interaction (ref.)						
Uncooperative	-2.15*** (.09)	-1.94*** (.08)	-2.26*** (.11)	-2.03*** (.10)	-2.27*** (.11)	-2.05*** (.10)
Prior interaction	1.03*** (.07)	0.95*** (.06)	1.13*** (.08)	0.98*** (.07)	1.16*** (.08)	0.97*** (.07)
Encapsulated interests	1.13*** (.06)	1.11*** (.06)	1.16*** (.08)	1.16*** (.07)	1.18*** (.08)	1.16*** (.07)
Goodwill	1.76*** (.07)	1.48*** (.06)	1.93*** (.08)	1.50*** (.07)	1.95*** (.08)	1.49*** (.07)
Virtuous dispositions	1.66*** (.07)	1.44*** (.06)	1.70*** (.08)	1.49*** (.07)	1.71*** (.08)	1.48*** (.07)
No contract (ref.)						
Non-binding contract	-0.04 (.04)	0.05 (.04)	-0.09 (.05)	0.08 (.05)	-0.07 (.05)	0.08 (.05)
Binding contract	0.62*** (.05)	0.28*** (.04)	0.58*** (.05)	0.32*** (.05)	0.61*** (.06)	0.32*** (.05)
No regulation (ref.)						
Non-monetary regulation	0.55*** (.05)	0.19*** (.04)	0.55*** (.06)	0.20*** (.05)	0.56*** (.06)	0.20*** (.05)
Monetary regulation	0.71*** (.25)	0.40*** (.04)	0.73*** (.06)	0.39*** (.05)	0.74*** (.06)	0.40*** (.05)
Constant	4.23*** (.39)	4.29*** (.39)	4.38*** (.45)	5.06*** (.38)	6.16*** (.71)	6.32*** (.77)
var(u_{0j})	0.78*** (.06)	0.87*** (.08)	1.00*** (.08)	0.98*** (.08)	0.92*** (.08)	0.95*** (.08)
var(e_{ij})	2.64*** (.08)	2.00*** (.07)	2.55*** (.10)	1.98*** (.09)	2.54*** (.10)	1.97*** (.09)
Other vignette dimensions	Yes	Yes	Yes	Yes	Yes	Yes
Vignette dummies	Yes	Yes	Yes	Yes	Yes	Yes
Individual-specific mean dim.	No	No	No	No	Yes	Yes
Individual-level covariates	Yes	Yes	Yes	Yes	Yes	Yes
Exclusions[1]	No	No	Yes	Yes	Yes	Yes
Observations	9361	8999	6268	6113	6223	6093
Individuals	986	945	635	618	630	616

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, † $p < 0.10$ (two-tailed)

Note: unstandardized slopes (robust standard errors in parentheses).

[1] Excludes failed screeners, partial completes, and shared IP addresses.

REFERENCES

Abascal, Maria. 2015. "Us and Them: Black-White Relations in the Wake of Hispanic Population Growth." *American Sociological Review* 80: 789-813.

Alexander, Cheryl S. and Henry J. Becker. 1978. "The Use of Vignettes in Survey Research." *Public Opinion Quarterly* 42: 93-104.

Auspurg, Katrin and Thomas Hinz. 2015. *Factorial Survey Experiments*. Los Angeles, CA: Sage.

Auspurg, Katrin and Annette Jäckle. 2015. "First Equals Most Important? Order Effects in Vignette-Based Measurement." *Sociological Methods & Research*, Forthcoming.

Berinsky, Adam J., Gregory S. Huber, and Gabriel S. Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk.". *Political Analysis* 20: 351-68.

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: 739-53.

Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling. 2011. "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?" *Perspectives on Psychological Science* 6: 3-5.

Bullock, John G., Donald P. Green, and Shang E. Ha. 2010. "Yes, But What's the Mechanism? (Don't Expect an Easy Answer)." *Journal of Personality and Social Psychology* 98: 550-558.

Camerer, Colin. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.

Chandler, Jesse, Pam Mueller, and Gabriele Paolacci. 2014. "Nonnaïveté among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers." *Behavior Research Methods* 46: 112-130.

Cochran, William G. and Gertrude Cox. 1957. *Experimental Design*. New York, NY: Wiley.

Dülmer, Hermann. 2015. "The Factorial Survey: Design Selection and its Impact on Reliability and Internal Validity." *Sociological Methods & Research*, Forthcoming.

Faia, Michael A. 1980. "The Vagaries of the Vignette World: A Comment on Alves and Rossi." *American Journal of Sociology* 85: 951-954.

Gerber, Alan S. and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York, NY: W. W. Norton & Company.

Gunst, Richard and Robert L. Mason. 1991. *How to Construct Fractional Factorial Experiments, Vol. 14. The Basic Reference in Quality Control: Statistical Techniques*. Milwaukee, WI: ASQC Quality Press.

Hainmueller, Jens, Dominik Hangartner, and Teppei Yamamoto. 2015. "Validating Vignette and Conjoint Survey Experiments Against Real-World Behavior." *Proceedings of the National Academy of Sciences* 112: 2395-2400.

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2014. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments." *Political Analysis* 22: 1-30.

Harrell, Ashley and Brent Simpson. 2016. "The Dynamics of Prosocial Leadership: Power and Influence in Collective Action Groups." *Social Forces* 94: 1283-1308.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81: 945-960.

Horton, John J., David G. Rand, and Richard J. Zeckhauser. 2011. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics* 14: 399-425.

Jasso, Guillermina. 2006. "Factorial Survey Methods for Studying Beliefs and Judgments." *Sociological Methods & Research* 34: 334-423.

Kam, Cindy D. and Elizabeth N. Simas. 2010. "Risk Orientations and Policy Frames." *Journal of Politics* 72: 381-396.

Kuwabara, Ko. 2015. "Do Reputation Systems Undermine Trust? Divergent Effects of Enforcement Type on Generalized Trust and Trustworthiness." *American Journal of Sociology* 120: 1390-1428.

Kuwabara, Ko and Oliver Sheldon. 2012. "Temporal Dynamics of Social Exchange and the Development of Solidarity: "Testing the Waters" Versus "Taking a Leap of Faith."" *Social Forces* 91: 253-273.

Mason, Winter and Siddharth Suri. 2012. "Conducting Behavioral Research on Amazon's Mechanical Turk." *Behavior Research Methods* 44: 1-23.

Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman, and Jeremy Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2: 109-138.

Mundlak, Yair. 1978. "On the Pooling of Time Series and Cross Section Data". *Econometrica* 46: 69-85.

Paolacci, Gabriele, Jesse Chandler, and Panos Ipeirotis. 2010. "Running Experiments on Amazon Mechanical Turk." *Judgment and Decision Making* 5: 411-19.

Rasinski, Kenneth A. 1989. "The Effect of Question Wording on Public Support for Government Spending." *Public Opinion Quarterly* 53: 388-394.

Rossi, Peter H. and Andy B. Anderson. 1982. ''The Factorial Survey Approach: an Introduction.'' Pp. 15-67 in *Measuring Social Judgments: The Factorial Survey Approach*, edited by P. H. Rossi and St. L. Nock. Beverly Hills, CA: Sage

Sauer, Carsten, Katrin Auspurg, Thomas Hinz, and Stefan Liebig. 2011. "The Application of Factorial Surveys in General Population Samples: The Effects of Respondent Age and Education on Response Times and Response Consistency." *Survey Research Methods* 5: 89-102.

Schunck, Reinhard. 2013. "Within and Between Estimates in Random-Effects Models: Advantages and Drawbacks of Correlated Random Effects and Hybrid Models." *The Stata Journal* 13: 65-76.

Shank, Daniel B. 2016. "Using Crowdsourcing Websites for Sociological Research: The Case of Amazon Mechanical Turk." *The American Sociologist* 47: 47-55.

Shapiro, Danielle, Jesse Chandler, and Pam A. Mueller. 2013. "Using Mechanical Turk to Study Clinical Populations." *Clinical Psychological Science* 1: 213-220.

Shariff, Azim F. and Ara Norenzayan. 2007. "God is Watching You." *Psychological Science* 18: 803-809.

Simpson, Brent, Ashley Harrell, and Robb Willer. 2013. "Hidden Paths from Morality to Cooperation: Moral Judgments Promote Trust and Trustworthiness." *Social Forces* 91: 1529-1548.

Suri, Siddharth and Duncan J. Watts. 2011. "Cooperation and Contagion in Web-Based, Networked Public Goods Experiments." *PLoS One* 6: e16836.

Tversky, Amos and Daniel Kahneman. 1981. "The Framing of Decisions and the Psychology of Choice." *Science* 211: 453-458.

Weinberg, Jill D., Jeremy Freese, and David McElhattan. 2014. "Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-Based and a Crowdsource-Recruited Sample." *Sociological Science* 1: 292-310.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Boston, MA: MIT Press.