Social Science Research 59 (2016) 68–82

Contents lists available at ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch

Using crowdsourced online experiments to study context-dependency of behavior

Marc Keuschnigg[*, 1], Felix Bader [1], Johannes Bracher

Department of Sociology, LMU Munich, Konradstrasse 6, 80801 Munich, Germany

ARTICLE INFO

Article history:
Received 21 July 2015
Received in revised form 26 January 2016
Accepted 5 April 2016
Available online 12 April 2016

Keywords:
Context effects
Cross-country equivalence
Dictator game
Mechanical Turk
Raising the stakes
Ultimatum game

ABSTRACT

We use *Mechanical Turk*'s diverse participant pool to conduct online bargaining games in India and the US. First, we assess internal validity of crowdsourced experimentation through variation of stakes ($0, $1, $4, and $10) in the Ultimatum and Dictator Game. For cross-country equivalence we adjust the stakes following differences in purchasing power. Our marginal totals correspond closely to laboratory findings. Monetary incentives induce more selfish behavior but, in line with most laboratory findings, the particular size of a positive stake appears irrelevant. Second, by transporting a homogeneous decision situation into various living conditions crowdsourced experimentation permits identification of context effects on elicited behavior. We explore context-dependency using session-level variation in participants' geographical location, regional affluence, and local social capital. Across "virtual pools" behavior varies in the range of stake effects. We argue that quasi-experimental variation of the characteristics people bring to the experimental situation is the key potential of crowdsourced online designs.

1. Introduction

The last five years have seen a rapid increase in crowdsourced recruiting for social science experiments over the Internet. Crowdsourcing platforms facilitate fast and inexpensive implementation of online experiments on large scales (Mason and Suri, 2011), tapping into real online labor markets (Horton et al., 2011) sustained by a substantially broader and more cross-national population (Berinsky et al., 2012) than in traditional experimental subject pools (see Henrich et al., 2010 for a prominent critique). Recently, however, some have questioned the quality of inferences from crowdsourced samples (Chandler et al., 2014; Gupta et al., 2014; Marder, 2015), particularly raising concerns regarding experience and self-selection of participants compromising online experiments' internal and external validity.

Against this backdrop we conducted a cross-national online experiment using Amazon's *Mechanical Turk* for recruitment ($N = 991$). We utilize international participation in a homogeneous online environment to circumvent common pitfalls of traditional cross-regional experimentation including multiple laboratories, ambiguous translations, and experimenter effects (Roth et al., 1991). We focus on fairness behavior among participants from the US and India using the Ultimatum Game (UG) and the Dictator Game (DG) for data generation. Both countries are quite distant on various axes of cultural differentiation (cf., Bonikowski, 2010; Huntington, 1993; Inglehart and Baker, 2000).

* Corresponding author. Present address: Institute for Analytical Sociology, Linköping University, 601 86 Norrköping, Sweden.
E-mail addresses: keuschnigg@lmu.de (M. Keuschnigg), felix.bader@lmu.de (F. Bader), jbracher@outlook.de (J. Bracher).
[1] M.K and F.B. contributed equally to this work.

http://dx.doi.org/10.1016/j.ssresearch.2016.04.014
0049-089X/

Our aim is twofold: First, we validate our set-up by experimental variation of monetary stakes ($0, $1, $4, and $10). We do not claim originality in this design for testing reproducibility of laboratory results. Others have posed similar research questions before (Amir et al., 2012; Raihani et al., 2013), and recent studies systematically compared results from crowdsourced online experiments and behavioral games in the laboratory (e.g. Hergueux and Jacquemet, 2015; Horton et al., 2011). We have extended prior research, however, enhancing functional equivalence in design (Harkness et al., 2003; Lynn, 2003) by adjusting stakes with respect to country-specific purchasing power.

Second, we will argue that crowdsourced online experiments provide a valuable complement to laboratory research transporting a homogeneous decision situation into various living conditions and social contexts. In fact, we believe that quasi-experimental variation of the characteristics people bring to the experimental situation is the key potential of crowdsourced online designs. Bringing context back into social experiments is particularly relevant for sociological research allowing inclusion of macro variables to explain individual behavior (Blalock, 1984; Coleman, 1990).

We exploit session-level variation in participants' geographical location to explore context-dependency of economic behavior. Our novel contribution thus follows from an analysis of what we call "virtual pools:" Crowdsourcing permits flexible grouping of participants along context variables (such as regional average income or social capital) to estimate effects of social strategies learned in participants' local environments. Virtual pools thus offer refined opportunities for comparison across social strata, regions, and cultures. To assess the importance of context our estimated stake effects provide a valuable benchmark.

Bargaining games concern the division of money between two players and model situations of conflict between self-interest and socially desirable behavior. UG and DG thus reveal expectations about valid social norms in a particular population (Bicchieri, 2006; Elster, 2007; Henrich et al., 2005).[2] We use bargaining games specifically because norms of fairness are strongly conditional on local context: Subjects' tendency to adhere depends on beliefs of others' compliance and others' normative expectations (Bicchieri, 2006; Bicchieri and Xiao, 2009).

The remainder proceeds as follows: First, we summarize strengths and weaknesses of global crowdsourcing platforms as tools for social experimentation (Section 2). We then outline our experimental design and hypotheses (Section 3). Section 4 introduces our approach of virtual pools. In Section 5 we evaluate the internal validity of crowdsourced online experiments, scrutinize potential influence of socio-demographics, and demonstrate exemplary analyses of virtual pools. Section 6 concludes with a discussion of challenges that lie ahead.

2. Crowdsourced experimentation

Crowdsourcing platforms aid implementation of large-scale online experiments (Berinsky et al., 2012; Buhrmeister et al., 2011; Mason and Suri, 2011). To this date Amazon's *Mechanical Turk* (MTurk) sustains the largest and most diverse participant pool. In 2014 MTurk's workforce exceeded 500,000 participants in 190 countries (Paolacci and Chandler, 2014), 60% from the US; India provides the largest portion of non-Americans, about one third of participants (Ross et al., 2010). Altogether, English serves as the commercial language.

MTurk has been launched in 2005 to provide clients (requesters) access to registered participants (workers) for hard-to-automatize "human intelligence tasks" (HITs). Requesters either upload tasks or redirect workers to outside websites. Each worker can choose from a continually updated list of HITs describing tasks, required qualifications, and payments. Following successful completion, requesters reward workers via MTurk's payment scheme (Amazon keeps an additional 10% of payments). MTurk allows configuration of HITs as surveys and experiments and so has proven increasingly useful for data generation in economics (e.g. Horton et al., 2011), political science (e.g. Berinsky et al., 2012), psychology (e.g. Crump et al., 2013), and sociology (e.g. Weinberg et al., 2014).[3]

MTurk is considered a real online labor market (Horton et al., 2011) in which workers seek profit-maximizing allocation of time and qualification. Unlike laboratory set-ups, Internet-based experimenting occurs in a relatively natural context limiting bias from unfamiliar testing conditions (Reips, 2002; Rhodes et al., 2003). Samples from MTurk are substantially more diverse in terms of participant characteristics than those from traditional experimental subject pools. American MTurk samples have repeatedly shown to be more representative of the US population than standard local convenience samples used in survey and laboratory research, but less representative than national survey samples and high-quality Internet panels (Berinsky et al., 2012; Buhrmeister et al., 2011). Compared to traditional laboratory research crowdsourced experimentation also reaps monetary benefits: Maintaining a physical laboratory becomes obsolete and show-up fees no longer need to cover participants' travel expenses.

Crowdsourced experimentation is not free from methodological challenges. Most importantly, researchers obtain no direct control over participants' surroundings. The upside is increased anonymity both among participants and towards the experimenter. Lack of control, however, permits larger variation in experimental conditions. For example, participants might be observed when taking the experiment or use the Internet to look up eligible answers and strategies. In principle, this is a

[2] Social norms are informal rules proscribing or prescribing certain behaviors. If perceived valid among a sufficiently large fraction of the population, and thus supported by mutual expectations, social norms generate regular patterns of behavior (Bicchieri, 2006; Bicchieri and Xiao, 2009).

[3] In June 2015 *Web of Science* listed 326 social science publications using MTurk for data generation. For technical implementation of experiments see Buhrmeister et al. (2011), Mason and Suri (2011), and, particularly, the blog *Experimental Turk* (2015). The platform is accessible at www.mturk.com.

threat to internal validity common to all online experiments (Reips, 2002; Rhodes et al., 2003). Similarly, attrition can be a problem as participants are more likely to drop out of anonymous online experiments than in-vivo sessions (Mason and Suri, 2011). Buhrmeister et al. (2011) argue that realistic compensation mitigates biases for short-duration tasks and generates data no less reliable than those obtained by traditional methods.

More specifically, MTurk workers also reportedly pay less attention to experimental instructions than do laboratory subjects (Goodman et al., 2012). The authors stress the importance of control questions to mitigate distraction and improve statistical power. Moreover, worker forums such as *Turkopticon* and *Turker Nation*, which rate requesters and HITs, may violate treatment integrity through cross-talk (Chandler et al., 2014). Projects discussed on forums, on the other hand, benefit from increased attention (Mason and Suri, 2011).[4]

Recently, some have criticized crowdsourced experimentation for relying on overly experienced participants (Chandler et al., 2014; Marder, 2015; Rand et al., 2014). Familiarity with social science theories and designs might jeopardize internal validity, but "the majority of MTurkers are not chronic study participants" (Berinsky et al., 2012, p. 365). Further, participants' full understanding of the decision situation is essential for economic games (as opposed to many psychological protocols).

Finally, valid cross-regional comparisons require homogeneous sampling. Self-selection into MTurk, however, varies strongly across countries. In India, for example, working for "American" MTurk is fairly respectable whereas in the US many consider it low-status work (Gupta et al., 2014). Low Internet penetration and language barriers surely contribute to Indian participants being considerably more affluent and better-educated than the overall population.

We address these concerns in both design and analysis. Redirecting workers to our own website enables drop-out analysis as we have data on HIT acceptance, stage of drop-out, and completion. We introduced control questions at every stage of the experiment and use homogeneous stakes in real terms. We minimize potential biases by controlling for self-administered information on participants' surroundings, experimental experience, and socio-demographics. Statistical control for regional self-selection is particularly important if one seeks to identify behavioral differences across virtual pools.

3. Design and hypotheses

In this section we spell out our experimental design, provide a rationale for our hypotheses, and summarize prior results on bargaining games.

3.1. Experimental games

To validate our set-up (variation of stakes) and demonstrate context-dependency of economic decision-making (virtual pools), we employ one-shot Ultimatum (UG) and Dictator Games (DG) in the US and India. In both decision situations participants have to choose between self-interested and socially desirable behaviors (Forsythe et al., 1994; Güth et al., 1982). Observed behavior provides an estimate of distributive norms because "fair" and profit-maximizing strategies diverge. Bargaining games thus reveal expectations about valid norms of fairness and informal sanctioning in specific populations (e.g. Bicchieri, 2006; Bicchieri and Xiao, 2009; Elster, 2007). We use bargaining games in particular because we hypothesize that such norms are strongly conditional on local context.

Laboratory experimental research has applied UG and DG heavily in measuring social preferences, i.e. individual motivations not restricted to one's own well-being but including the well-being of others (Camerer and Fehr, 2004). Marginal totals are well-documented for various populations (Camerer, 2003; Cardenas and Carpenter, 2008; Henrich et al., 2010) and stake levels (Camerer and Hogarth, 1999; Carpenter et al., 2005). Each protocol is easily understandable by participants of different social backgrounds and training.

> **Ultimatum Game**. A proposer receives a stake and can decide on how much of the pie (0–100%) she offers to a responder. The responder can accept her offer (and both players receive their share); the responder can decline her offer (and both players receive nothing).

Assuming rationality, common knowledge of rationality, and selfish preferences standard prediction follows from backward induction: Being certain that a responder will prefer any positive offer to a zero-payoff, a proposer will maximize her income by offering the smallest possible positive amount—which the responder accepts.

Apparently, experimental research has shown that real human behavior is not purely profit-maximizing. In a meta-analysis of 75 results Oosterbeek et al. (2004) report mean offers of 40.4%. In most countries modal offers range between 40% and 50% of the endowment, and responders reject offers below 20% about half of the time (Camerer, 2003). It is noteworthy, however, that observed behavior among some populations fits the standard prediction rather well suggesting that strategy selection is context-dependent (Henrich, 2000).

To explain giving in bargaining games social scientists proposed various motivations including altruism and warm glow (Andreoni, 1990), fairness (Bicchieri, 2006), inequality aversion (Fehr and Schmidt, 1999), reciprocity (Diekmann, 2004),

[4] We continuously checked *Turkopticon*, *Turker Nation*, and a number of smaller forums, and found no indication anyone was discussing our research purpose. Participants instead commended our HITs for above average compensation.

identity (Akerlof and Kranton, 2000), and social desirability (Franzen and Pointner, 2012). These approaches assume that people adhere to intrinsically motivated and socially sanctioned rules of behavior, perceive the profit-maximizing strategy as antisocial, or enjoy the act of giving for their self-image. In either case, non-compliance with social norms bears significant costs for individual decision-makers (e.g. Elster, 2007; Fehr and Fischbacher, 2004).

Still, equitable offers in the UG might follow from purely strategic reasoning such that proposers fear responders' rejection of offers perceived as "unfair" (e.g. Camerer and Fehr, 2004; Wells and Rand, 2013). Indeed, responders from both student pools and non-standard samples frequently decline low offers (Camerer, 2003; Cardenas and Carpenter, 2008). We interpret rejection as informal sanctioning of norm violations: Responders' behavior indicates willingness to reciprocate negatively giving up their own profits to enforce normative principles (Camerer and Fehr, 2004; Rauhut and Winter 2010).

We discern social preferences (including social desirability towards the experimenter) from strategic fairness by comparing offers in UG and DG.

Dictator Game. A dictator receives a stake and can decide on how much of the pie (0–100%) she passes to a receiver.

DG features a parametric situation without direct interaction and thus fully removes fear-of-sanction. Standard prediction suggests that dictators keep 100% of the endowment. Giving can only stem from social preferences. One-shot DG should thus provide a more realistic estimate of social motivations. A meta-analysis of more than 600 DG results reports a mean allocation of 28.4% (Engel, 2011). We expect:

H1 Average offer in DG will be substantially lower than in UG.

H2 H1 will hold only under positive stakes. When allocation is hypothetical (and fairness costless) we expect no difference in average offer across UG and DG.

In our demonstration of virtual pools we will particularly make use of DG's "weak" design properties in that average allocation can change intensely depending on subject pool and context (Camerer and Fehr, 2004; Hoffman et al., 1996). In this sense, we expect that behavior in a weak decision situation reflects not only optimal response to institutional conditions but also social norms and beliefs about their salience within a social environment.

3.2. Monetary stakes

Experimentalists typically assume that participants will more likely adhere to social norms in hypothetical allocation games than when real money is at stake (Camerer and Hogarth, 1999; Heyman and Ariely, 2004). As the costs of both fairness and informal sanctioning increase we should find less socially desirable behavior in high-stakes decisions. Validating our online design we test whether monetary incentives undermine the power of social norms.

Potential stake effects are important for experimental methodology and are well-documented over the last 20 years (e.g. Carpenter et al., 2005; Forsythe et al., 1994; Hoffman et al., 1996): Stakes induce involvement and effort in memory and learning tasks but are hardly a leverage in economic games. "When behavior does change, incentives can be interpreted as shifting behavior away from an overly socially-desirable presentation of oneself to a more realistic one" (Camerer and Hogarth, 1999, p. 8).

Typically, raising the stakes increases selfish behavior in the DG (e.g. Cardenas and Carpenter, 2008; Forsythe et al., 1994; Sefton, 1992) but does not affect proposer behavior in the UG (e.g. Cameron, 1999; Carpenter et al., 2005; Hoffman et al., 1996). UG responders, on the other hand, reduce their percentage threshold for acceptable offers when endowments are greater. Hence, rejection rates decrease with rising stakes (e.g. Forsythe et al., 1994; Munier and Zaharia, 2002; Slonim and Roth, 1998). Most studies raise stakes from $5 to $10 (Forsythe et al., 1994) or from $10 to $100 (Carpenter et al., 2005; Hoffman et al., 1996). Others induce much larger incentives (Andersen et al., 2011; Cameron, 1999; Slonim and Roth, 1998).

We manipulated stakes in a 2 × 4 design (Table 1) and tested stake effects by between-subject comparison. To achieve functional equivalence (Harkness et al., 2003) we weighted payoffs for Indian participants using a purchasing power parity (PPP) conversion factor of 0.4 (OECD, 2013): Participants from India decided, for example, on a nominal allocation of $1.6 while US participants had to split $4. We used tokens to further homogenize decision situations across countries (Roth et al., 1991): A first-mover received 10 tokens, each representing 10% of the stake. In the $0-treatment participants decided on a hypothetical split of 10 tokens representing "play money." We allowed individual offers only to be multiples of 1 (including 0).

Table 1
Variation of stakes and observations per treatment.

USA				India			
Stake	DG	UG-P	UG-R	Stake	DG	UG-P	UG-R
$0	61	61	62	$0	61	62	61
$1	61	61	61	$0.4	62	61	62
$4	68	64	64	$1.6	67	64	64
$10	61	61	61	$4	61	61	61

DG: dictators; UG-P: proposers; UG-R: responders.

Our design reveals whether Indian and US participants show similar behavior once stakes are equal in real terms. For consistency with laboratory results we expect:

H3 Raising the stakes does not reduce the average offer in UG as fear-of-sanction drives behavior. Offers in DG, instead, reflect the degree of self-interest and thus respond to stake manipulation. Similarly, UG responders' willingness to punish unfair offers decreases with stake size.

We are particularly interested in the effect of introducing monetary stakes altogether. Most likely it is the change from $0 to PPP$1 that makes a behavioral difference, shifting outcomes towards the standard prediction (cf., Carpenter et al., 2005):

H4 Raising stakes from $0 to PPP$1 has the largest marginal effect on both dictator (DG) and responder behavior (UG).

Two prior studies explicitly tested for stake effects in international samples recruited at MTurk. Amir et al. (2012) conducted DG, UG, Trust Game, and Public Goods Game offering a $0.4 show-up fee and stakes of $0 or $1. Their results indicate that even at $1-stakes crowdsourced online experiments generate marginal totals comparable to laboratory results. Raihani et al. (2013) varied stakes in the DG ($1, $5, and $10), differentiating between US and Indian participants. Indians appeared less generous at higher stakes, a pattern not consistent with US participants. Neglecting purchasing power differences, however, both studies lack cross-national comparability.

3.3. Experimental procedure

Data generation included recruiting, randomization, instructions, eliciting behavioral data, collection of survey data, and payoff.

Recruiting. We conducted our experiment from September 4 to December 13, 2014 using MTurk for recruiting. Our HIT read "Participate in decision experiment: Work on three simple decision tasks and answer a short questionnaire" offering a $0.5 reward and a potential bonus "depending on your and other players' decisions." One-time-only participation was restricted to US American and Indian workers. For each participant pool we posted two HITs daily, one in the early morning and one in the late afternoon (local time). For each daily session we recruited as many US Americans as we had recruited Indians earlier that day. Thus, our sample consists (almost) equally of US Americans and Indians (see Table 1) and exhibits session-level variation in both individual and population characteristics.

Randomization. Upon acceptance we redirected workers to our own website, randomly assigning each arrival to a specific stake level, sequence of games, and first- and second-mover roles. We instructed participants to be randomly matched to a perfect stranger in each game. To avoid waiting time (and drop-out), actual matching, however, occurred only before payoff: We randomly selected one of three games by each finalist to pair with a decision randomly drawn from the pool of preceding participants (without replacement).

Instructions. Our written instructions (see Supplementary material) map participants' choices to payoffs as clearly as possible using numerical examples but avoiding suggestion of specific strategies or frames. We refrained from multi-language instructions using simple English common to almost all MTurk HITs. Monitoring attentiveness and understanding of tasks is crucial to maintaining statistical power, particularly in crowdsourced experiments (Goodman et al., 2012). To improve data quality, we included a set of control questions each participant had to answer correctly before making the decision in each game (we allowed for three trials per game). Additionally, we included the Prisoner's Dilemma (PD) as a "control game" featuring more complex instructions than the simple bargaining games. Our analyses include only participants who understood PD instructions correctly.[5]

Behavioral data. Each arrival participated in three 2-person games, UG, DG, and PD, with random ordering and role assignments at a random stake level conditional on country of residence. The absence of feedback in-between games secured independence of sequential behavior.[6] To elicit UG responders' minimal acceptable offer (MAO), we employed the strategy method asking responders to value all feasible offers (Rauhut and Winter 2010). We refrained from using the standard measure of responder behavior (second player's response to proposer's actual offer) because most offers exceeded responders' MAO (censoring the response variable).

Survey data. To finalize the HIT we requested participants to fill out a 2-page questionnaire including items on sociodemographics, personal experience with experimental games, geographical location at state level, and both physical and social surroundings during participation (see section 4 for a variable list). We administered the questionnaire at the end of the experiment to minimize respondents' motivation to misreport in order to increase their chance for participation.

Payoff. We restricted payment to fully completed HITs ($N = 991$). We rejected submissions which failed to provide a correct answer to at least one control question in UG, DG, and PD (577 subjects) or otherwise dropped out before completion (232 subjects).[7] Payoff included the show-up fee ($0.5) and a $0.5 bonus for completing the questionnaire (for participants in the $0 treatment) or a variable bonus computed from ego's (and partner's) decision in one of three games (for participants in

[5] This study will not consider PD results.

[6] Controlling for the sequence of games, it shows that decisions in UG and DG are independent of having participated in the PD before or not.

[7] Drop-out is independent of whether we offer hypothetical or real stakes ($\chi^2 = 0.303$; $p = 0.582$; $N = 1{,}800$), securing internal validity of estimated treatment effects.

positive stake treatments). We made randomized rewards (Bolle, 1990) common knowledge in our instructions, explaining that each game would determine individual payoffs with probability 1/3. On average US (Indian) participants earned $2.34 ($1.27) in 14 min. The difference in nominal dollar payoffs is due to stakes' purchasing power adjustment. Participants received a completion code together with feedback on the disbursed game. Payoffs were transferred via MTurk within 24 h after participation.

Our set-up conforms to ethical guidelines (e.g. Williamson, 2014): Each subject received a detailed HIT description before accepting participation (informed consent). Participation was anonymous. Each participant had equal probability of receiving a high-stake treatment. Average PPP-adjusted payoff was above US minimum wage. The study used no deception apart from delayed matching to generate payoffs.

4. Virtual pools

Apart from the well-known strengths of laboratory designs for causal inference, local confinement of samples and results has motivated questions as to the external validity of social science experiments (e.g. Henrich et al., 2010; Levitt and List, 2007; Roth et al., 1991). Crowdsourced online experiments provide a valuable complement allowing flexible grouping of participants along contextual data. We explore context-dependency of fairness behavior using session-level variation in participants' geographical location. Estimation of context effects aims at explaining individual behavior by combining individual- and group-level regressors (Blalock, 1984; Blau, 1960).

Studies of virtual pools first mirror standard sub-group analyses contrasting participants from different regions. More importantly, however, participants' geographical locations provide an interface for direct inclusion of regional macro variables potentially influencing individual behavior. This capability of crowdsourced experimentation is particularly relevant for sociological research which—unlike most experimental research in economics and psychology—fully acknowledges the importance of context effects in a multi-level explanation of individual action (Coleman, 1990; Hedström and Swedberg, 1998; Merton, 1949).

The last 15 years have seen a sharp increase in experiments conducted at multiple locations including developing countries and small-scale societies (see Cardenas and Carpenter, 2008 for an overview). Above all Henrich et al. (2005, 2010) have added to our understanding of culturally molded decision-making and its interaction with socio-economic development. However, cross-regional comparisons for estimating effects of living conditions and cultural patterns on economic behavior have run into obstacles due to limited transferability of standardized decision situations into parallel laboratory setups (e.g. Baumard and Sperber, 2010; Cardenas and Carpenter, 2008). Our online design circumvents these methodological challenges by transporting a homogeneous decision situation into variable living conditions and social contexts.

Still, individual attributes and locally molded strategies evade randomization and enter the online laboratory as quasi-experimental conditions (Shadish et al., 2001). In this connection, region-specific self-selection into participation aggravates cross-regional comparison even in a functionally equivalent design. Hence, homogeneity assumptions regarding comparability of virtual pools are essential (Blalock, 1984).

4.1. Socio-demographics

Participants provided information on personal characteristics and surroundings during experimentation (Table 2). Compared to traditional subject pools our crowdsourced sample is considerably more diverse, including mostly non-student (87%) and fully employed (48%) participants. Mean age is 32 years and 38% have children (see Table A6 in the Supplementary material for a pooled description). Still, our sample is hardly representative of the general population in the US and especially in India, most pronouncedly in that our sample features a larger share of men, university graduates, and the fully-employed

Table 2
Socio-demographics.

		USA		India		Diff.
Construct	Variable	Mean	SD	Mean	SD	*p*
Individual income	PPP-adjusted per capita equivalent household income	2430.39	6826.68	1015.49	8109.70	0.003
Experience	number of prior experiments	21.82	32.22	3.23	10.88	<0.001
Age	age in years	33.45	10.44	31.03	9.53	<0.001
Religiosity	self-assessment (1–10)	3.22	2.96	6.40	2.74	<0.001
Gender	male (0,1)	0.56		0.71		<0.001
Minority	Non-White; Non-Hindu (0,1)	0.21		0.29		0.003
Education	university degree (0,1)	0.53		0.88		<0.001
Employment	full-time employed (0,1)	0.48		0.48		0.974
Student	currently student (0,1)	0.12		0.14		0.305
Children	respondent has ≥1 child (0,1)	0.32		0.44		<0.001
Social surrounding	participation observed (0,1)	0.01		0.04		0.002
Physical surround.	participation not at home (0,1)	0.10		0.11		0.684

We report *p*-values from *t*- and χ^2-tests for between-country differences.

than does the general population. The sample from India is particularly biased towards subjects from smaller, better-educated, and relatively affluent households (cf., Gupta et al., 2014).

We are particularly interested in potential effects of personal income and experimental experience (both vary considerably within and between country samples). First, one could hypothesize that individual affluence reduces selfishness: Participants less dependent on income generated at MTurk might be more generous in bargaining games.

Second, given recent criticism of MTurk's overly experienced work force (Chandler et al., 2014; Marder, 2015) procedural knowledge about experimental games, or nonnaïveté, might threaten internal validity. More specifically, estimation of context effects relies on participants' resorting to regionally imprinted rules of behavior in "weak" experimental situations. In this connection, Rand et al. (2014) proposed what they call the "social heuristics hypothesis:" One can expect inexperienced participants to rely on successful strategies learned in daily interaction; frequent participation in social experiments, however, provides the opportunity to calibrate one's impulsive responses and to arrive at behavior more consistent with the standard theory (see also Levitt and List, 2007).[8] If experience alters elicited behavior local context no longer defines the socially acceptable, cognitively familiar, and habitual scope of action for nonnaïve subjects. In this case, one could only estimate context effects from data provided by inexperienced participants.

4.2. Context data

Our application of virtual pools tests for behavioral variation between US and Indian participants and state-wise differences within the US. We also tested context-dependency within India using available indicators. Due to strong self-selection among Indian participants and lower quality of context data, however, we only briefly discuss these results in Subsection 5.2.

Following conventional test strategy (Cardenas and Carpenter, 2008) we interpret behavioral differences across virtual pools only given two criteria: First, regional differences remain statistically significant after controlling for individual demographics. Second, to speak of cross-country differences, between-country variation needs to be larger than within-country differences along some salient (cultural) fault line.

We pose cross-country differences against a North–South comparison within the US introducing a dummy "South" for the eleven former Confederate states (see Table 3). We chose this benchmark because the "South is still singled out as the most peculiar of the American regions" (Marsden et al., 1982, p. 1023). The roots of this cultural divide have provoked discussion

Table 3
Context variables across US states.

State	*n*	South	Average income	Social capital	State	*n*	South	Average income	Social capital
Alabama	4	1	37,493	−1.07	Nebraska	2	0	**47,073**	**1.15**
Arizona	14	0	37,895	**0.06**	Nevada	8	0	40,077	−1.43
Arkansas	5	1	37,751	−0.50	New Hampshire	2	0	**53,149**	**0.77**
California	53	0	**50,109**	**−0.18**	New Jersey	15	0	**56,807**	−0.40
Colorado	9	0	**48,730**	**0.41**	New York	23	0	**56,231**	−0.36
Connecticut	7	0	**62,467**	**0.27**	North Carolina	22	1	39,646	−0.82
Delaware	3	0	**45,942**	**−0.01**	Ohio	23	0	42,571	**−0.18**
Florida	43	1	42,645	−0.47	Oklahoma	3	0	43,138	**−0.16**
Georgia	14	1	39,097	−1.15	Oregon	9	0	41,681	**0.57**
Hawaii	3	0	**46,396**	–	Pennsylvania	37	0	**47,727**	**−0.19**
Illinois	15	0	**48,120**	−0.22	Rhode Island	3	0	**48,838**	**−0.06**
Indiana	12	0	39,433	**−0.08**	South Carolina	8	1	36,934	−0.88
Iowa	1	0	45,115	**0.98**	Tennessee	11	1	40,654	−0.96
Kansas	2	0	**45,546**	**0.38**	Texas	27	1	45,426	−0.55
Kentucky	10	0	37,654	−0.79	Utah	1	0	37,766	**0.50**
Louisiana	4	1	42,287	−0.99	Vermont	2	0	**47,330**	**1.42**
Maine	3	0	42,071	**0.53**	Virginia	16	1	**49,710**	−0.32
Maryland	17	0	**55,143**	−0.26	Washington	10	0	**49,583**	**0.65**
Massachusetts	11	0	**59,182**	**0.22**	Washington, D.C.	1	0	**76,532**	–
Michigan	15	0	40,556	**0.00**	West Virginia	5	0	36,644	−0.83
Minnesota	7	0	**48,711**	**1.32**	Wisconsin	5	0	44,585	**0.59**
Mississippi	2	1	34,333	−1.17	Median	8		45,426	−0.22
Missouri	7	0	41,613	0.10	Mean	11	0.32	46,082	−0.26
Montana	1	0	40,601	1.29	SD	11.35		6,366	0.50

We had no participants from Alaska, Idaho, New Mexico, North Dakota, South Dakota, Wyoming ($n = 0$). We report sample median, mean, and SD (weighted by n) and display above-median values of context variables in bold face.

[8] For a similar argument see Bicchieri (2006, p. 5): "[W]hen faced with a new situation, we immediately search for cues about how to interpret it or what is appropriate behavior for that situation. It is conjectured that we compare the situation we face with others we remember that possess similar characteristics, and that this comparison activates behavior that is considered most 'normal' for this type of situation."

since Tocqueville (1840). Several research areas point out diverging normative imprinting in America's South (see, e.g. Lee et al., 2007; Marsden et al., 1982; Nisbett and Cohen, 1996), particularly stressing a Southern "culture of honor, in which even small disputes become contests for reputation and social status" (Cohen et al., 1996, p. 945).

Further, we group US states along context variables. We expect norm adherence to vary with regional socio-economic conditions (Bicchieri, 2006; Henrich et al., 2005). Selection of context variables requires theoretical guidance (an important issue we elaborate further in the concluding section). In order to secure internal validity and avoid "fishing" for effects (Shadish et al., 2001, p. 48) we limit inclusion of context data to two broadly reflected indicators, economic prosperity and social capital. Still, we cannot reject ecological fallacy (Robinson, 1950), nor do we test for underlying causal mechanisms (Hedström and Swedberg, 1998; Morgan and Winship, 2007, ch. 8). Instead, we report correlations between macro variables and individual choices in our experiment. Our demonstration thus remains exploratory.

Most social scientists agree that local practices are sensitive to economic conditions. Following this tradition, we assume that, *ceteris paribus*, affluent communities maintain pro-social norms more easily, as their members depend less on short-term benefits of norm violation. Normative repudiation of self-serving strategies might, then, be lower among deprived communities. Material deprivation affects time and risk preferences and thus should also shape patterns of economic interaction on the macro level (e.g. Bardhan and Udry, 1999; Duflo, 2007).[9] To model variation in participants' regional economic conditions we match state-level yearly average personal income (US Bureau of Economic Analysis, 2015) to our behavioral data (see Table 3).

Besides economic conditions we expect fairness behavior to vary with social capital, i.e. local arrangements of institutions, social relations, and shared norms (e.g. Dasgupta and Serageldin, 2000; Lin, 2001). For our purpose we refer to social capital as a macro-level variable not assigned to individuals (as in, e.g. Bourdieu, 1984; Granovetter, 1973) but to regional populations (e.g. Fukuyama, 1995; Putnam, 1993). Our interest lies in sharing of social norms and maintenance of social control brought about by social coherence, civic engagement, and closely-knit networks within a local population. Following Coleman (1990) and others (e.g. Ostrom and Ahn, 2003; Sampson et al., 2002) we assume that social capital increases predictability of social interaction and affects beliefs on others' willingness to punish norm violation.

To evaluate whether local social capital translates to the online laboratory we include Putnam's (2000) state-wise index of social capital. This well-established indicator includes survey variables regarding (a) social activities such as meeting friends and attending club meetings, (b) civic engagement such as volunteering and committee membership, and (c) generalized trust as well as aggregate data regarding (d) density of civic organizations and voter turnout (see Putnam, 2000, p. 291 for description). We expect participants' decisions whether to violate distributive norms to depend on the extent of social capital in their area.

It scarcely surprises that state-wise average income and social capital positively associate ($r = 0.354$; $p = 0.020$; $N = 43$). We dichotomize context variables at their sample median (weighted by n). First, binary measures spare us to define a particular functional form of the relationship between context data and bargaining behavior (for which no theory exists). Second, our rough approximation of geographical location neglects variation of context variables within states. At least to some degree, binary measures account for local variation, as they merely group but do not rank (as in ordinal measures) or rate (as in cardinal measures) virtual pools along context values. Similarly, small n in some states becomes less of a problem given grouping.

5. Results

We first present experimental results on the interplay of UG and DG and describe sanctioning behavior at different stake size. Controlling for individual characteristics, we then provide suggestive evidence on context-dependency of bargaining behavior.

5.1. Experimental results

Fig. 1 displays average offers in UG and DG at varying stakes. The black bars in panel (a) represent social preferences—as measured in DG—whereas white extensions indicate good will due to strategic fairness among UG proposers (UG-P). Consistent with H1, offers in UG clearly exceed DG allocations. Across stake levels average offer is 4.0 in UG, whereas mean allocation in DG is 2.8 ($t = 8.991$; $p < 0.001$; $N = 997$).[10] Hence, results from our crowdsourced online experiment closely correspond to laboratory findings reporting an average offer of 4.04 in UG and 2.84 in DG (Engel, 2011; Oosterbeek et al., 2004).

[9] Cardenas and Carpenter (2008) report negative correlations between experimental measures of trust and country-level poverty, income inequality, and unemployment. Henrich et al. (2005) report greater pro-sociality among UG proposers stemming from economically integrated small-scale societies as opposed to hunter-gatherers without market embeddedness. Both studies, however, use data from multiple physical laboratories.

[10] We report α-errors of two-sided t-tests. We estimate differences between UG-P and DG from pooled data including both between- and within-subject variation. 50% of participants provide one observation each—either as first-mover in UG or in DG (between-subject variation). The other half provides two observations each—both as first-mover in UG and in DG (within-subject variation). Our findings are robust to sample limitations either to pure within-subject or pure between-subject comparison.

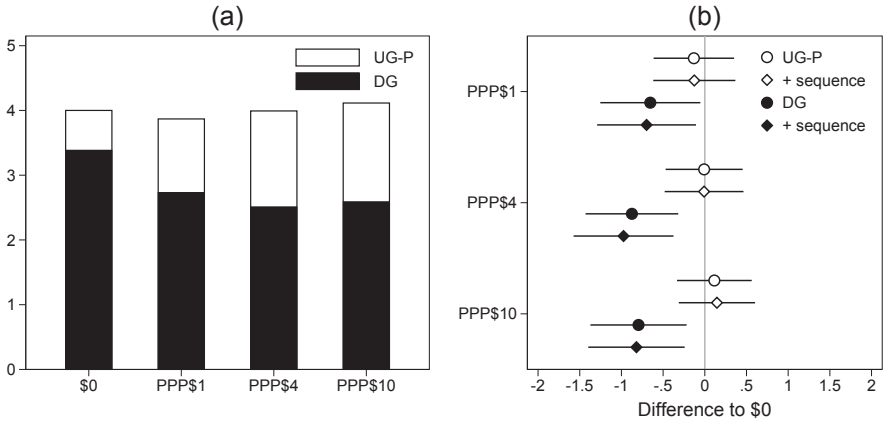

Fig. 1. Average offers in DG and UG at different stakes.

Consistent with H2, differences across UG and DG are smallest under hypothetical stakes; mean offers are 4.0 (UG) and 3.4 (DG). Hence, when fairness is costless dictators' allocations converge to levels almost as high as in UG (differences, however, remain significant; $t = 2.286$; $p = 0.023$; $N = 245$). Outcomes diverge under positive stakes: Across stake levels PPP$1, PPP$4, and PPP$10, UG proposers still offer 4.0 on average; in DG mean allocation drops to 2.6 tokens or 26% of the endowment ($t = 9.096$; $p < 0.001$; $N = 752$).

In panel (b) we test for significance of stake effects. The vertical zero-line represents average offer under hypothetical stakes: 4.0 in UG and 3.4 in DG. Circles represent unconditional means in UG (white) and DG (black); we have added 95% confidence intervals. Controlling for potential sequence effects (random ordering of games resulted in 6 possible sequences which we interacted with the role in the preceding game) we include a conditional mean per treatment (diamonds). Validating our design, results remain robust.

Our finding conforms to H3 and is thus fully in line with prior laboratory research (Carpenter et al., 2005; Hoffman et al., 1996): Raising the stakes reduces average offer in DG but not in UG. We find a similar pattern in both India and the US (results not plotted). Inducing monetary stakes reduces average allocation in DG from 3.5 to 2.6 tokens in India ($t = 3.009$; $p = 0.003$; $N = 251$) and from 3.3 to 2.6 tokens in the US ($t = 1.676$; $p = 0.095$; $N = 251$). In UG-P incentivizing remains ineffective in both India (3.5 vs. 3.8; $t = 0.840$; $p = 0.402$; $N = 248$) and the US (4.5 vs. 4.2; $t = 1.079$; $p = 0.282$; $N = 247$). If fear-of-sanction drives behavior, stake manipulations will not induce more selfish behavior. This result indicates the stronger design properties of UG compared to DG.

We also find support for H4, according to which raising stakes from $0 to PPP$1 should have the largest marginal effect on dictator behavior. Consistently, DG offers under positive stakes differ significantly from hypothetical allocations by 0.7 (PPP$1), 0.9 (PPP$4), and 0.8 tokens (PPP$10) whereas differences across positive stakes are statistically indiscernible. Hence, for online experimentation, monetary incentives are important; in line with most laboratory findings, however, stake size appears irrelevant.

Fig. 2 describes responder behavior.[11] In accordance with laboratory findings rejection rate sharply declines with increasing offers (panel (a)). On average the minimal acceptable offer (MAO) is 2.8, and 56% of responders would decline offers of 20% or less. Rejections of 1–3 tokens are more frequent at stakes $0 and PPP$1 when punishing is cheap. Increasing stakes reduce the percentage threshold of acceptable offers (panel (b)): In a hypothetical decision responders minimally accept 3.3 on average whereas they content themselves with 2.6 under positive stakes ($t = 3.681$; $p < 0.001$; $N = 464$). Again, and consistent with H4, increasing stakes from $0 to PPP$1 has the largest marginal effect across both countries. The decrease in aspirations appears similar in India (3.3 vs. 2.5; $t = 2.388$; $p = 0.018$; $N = 228$) and in the US (3.4 vs. 2.7; $t = 2.796$; $p = 0.006$; $N = 236$).

Demonstrating the validity of online experimentation our results are closely in line with prior laboratory findings. Moreover, as online subjects show consistent and expectable behavior they appear to have diligently read and comprehended the instructions provided.

5.2. Quasi-experimental results

To evaluate comparability of virtual pools, we first test the null hypothesis of no behavioral differences along socio-demographics. We then demonstrate context-dependency of bargaining behavior.

[11] We excluded 32 subjects with inconsistent strategy tables.

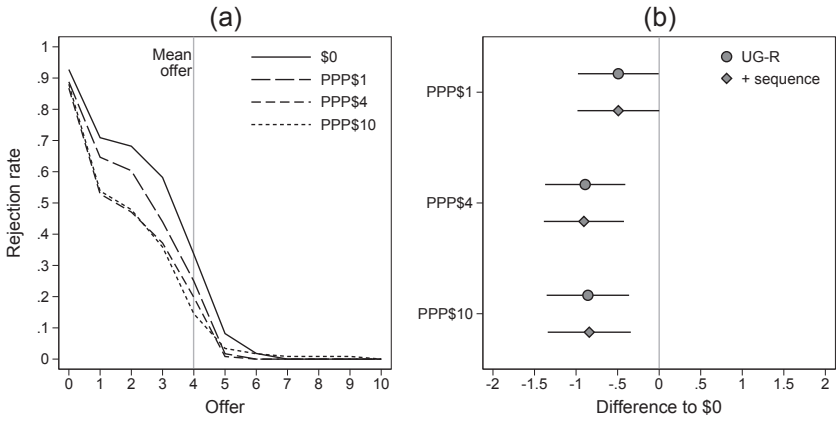

Fig. 2. Rejection rate and minimal acceptable offers in UG at different stakes.

5.2.1. Socio-demographic effects

In Table 4 we evaluate the influence of individual characteristics on elicited behavior. The constant represents the average offer (in models 1 and 2) or the minimal acceptable offer (in model 3) under hypothetical stakes for non-minority, non-student US females who participated unobserved at home, hold mean experience, income, religiosity, and age, and have no university degree, full-time job, or children. All other variables' estimates represent deviations from this sub-group's average behavior. To account for the multi-level data structure we use robust standard errors throughout.[12]

First, our experimental findings are robust against inclusion of socio-demographics. Conditioning on individual characteristics does not affect estimated treatment effects. Second, neither personal income nor experience relate to behavior regardless of the game or the role played (we use both skewed variables on a logarithmic scale).[13] Therefore, we will be able to estimate context effects using data from both naïve and nonnaïve subjects. Third, our measures of experimental surroundings do not affect elicited behavior. Fourth, there are apparent differences across games (and roles) regarding the potential for

Table 4
Differences across socio-demographics.

	DG offer		UG-P offer		UG-R MAO	
	(1)		(2)		(3)	
	β	p	β	p	β	p
Constant	4.046		4.532		3.612	
PPP$1	−0.667	0.024*	−0.077	0.756	−0.500	0.050
PPP$4	−0.922	0.002**	0.051	0.837	−0.893	0.000***
PPP$10	−0.767	0.010*	0.197	0.397	−0.843	0.001**
log Individual income	0.018	0.801	−0.006	0.901	−0.031	0.625
log Experience	−0.052	0.498	0.075	0.219	−0.065	0.351
Age	−0.002	0.885	−0.002	0.791	−0.008	0.501
Religiosity	0.103	0.007**	0.058	0.068	0.016	0.639
Male	−0.249	0.277	0.047	0.788	0.014	0.941
Minority	−0.154	0.515	−0.330	0.136	0.133	0.511
University degree	−0.399	0.111	−0.263	0.182	0.208	0.346
Full-time employed	−0.047	0.841	−0.084	0.654	−0.175	0.391
Student	0.301	0.397	0.014	0.962	−0.074	0.809
Children	0.504	0.045*	0.083	0.680	−0.303	0.170
Participation observed	−0.604	0.389	0.621	0.234	−0.054	0.918
Participation not at home	−0.272	0.450	0.487	0.118	−0.229	0.464
India	−0.312	0.283	−0.626	0.007**	−0.384	0.109
N	502		495		464	
R^2	0.091		0.061		0.073	

OLS regressions with sequence fixed effects. $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

[12] Hierarchical linear models with random intercepts on the country- or state-level yield results similar to these in Tables 4 and 5, and Fig. 3. Further, we re-parameterized our analysis to include Tobit regression (for left-censored dependent variables) and zero-inflated negative binomial regression (for count data) which produced nearly identical results. We stick to simple OLS.

[13] We also get null effects using linear or binary (median-split) specifications.

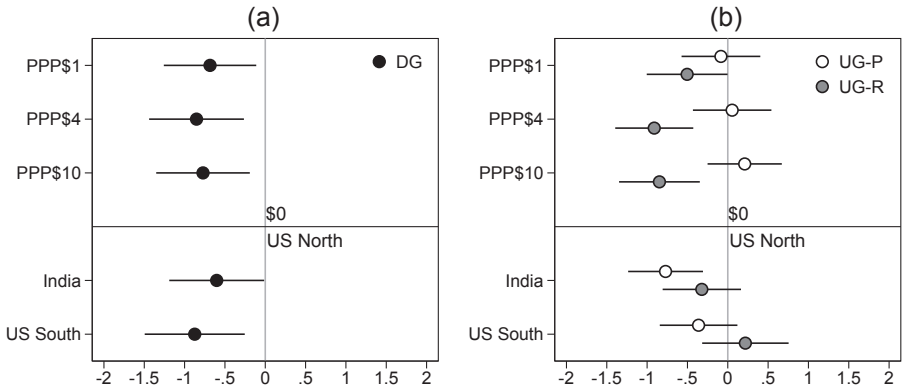

Fig. 3. Differences in DG and UG across countries and regions.

influence due to personal attributes. This corroborates our assumptions regarding the varied "weakness" of decisive situations: Dictator behavior appears most amenable to socio-demographic influence (model 1). In DG both religiosity and having children go along with increased social preferences.[14] In terms of standardized coefficients, however, monetary stakes matter more than demographics. MAO by UG responders, in turn, remains completely unaffected by individual characteristics (model 3). In UG we find a substantial country effect, with Indians offering on average 0.6 tokens fewer than Americans (model 2). We will substantiate this potential cultural difference below.

5.2.2. Context effects

In Fig. 3 we contrast behavior elicited in India and the US. To distinguish cross-country differences from ordinary regional variation we pitch the "India effect" against a North-South comparison within the US. In the upper part of both panels we report stake effects as a benchmark. Here, the vertical zero-line represents average behavior under hypothetical stakes. In the lower parts the zero-line indicates average behavior among US-Americans from "Northern" states (across all stake levels). For comparability of virtual pools we report conditional means under full control of socio-demographics (as in Table 4).

Panel (a) displays geographical variation in dictators' mean allocation. Both location dummies show significant effects of roughly the same size as our experimental treatments. On average Indian dictators offer 0.6 tokens fewer than participants from America's North. Unlike in Table 4 (model 1), where we contrasted Indian to US allocations, this difference is significant ($p = 0.045$). However, regional US variation appears larger (−0.9 tokens; $p = 0.006$). Given greater regional variation within the US, we must not interpret disparities between India and the US in the parametric DG as due to cross-country or even cultural differences.

Panel (b) displays a similar analysis for strategic UG. In line with Table 4 (models 2 and 3) we find a considerable cross-country difference for proposers (hollow circles) and an insignificant effect for responders (shaded circles). In both roles, absolute differences across countries are larger than variation within the US, indicating substantial country differences in ultimatum bargaining. For an explanation of proposer behavior, geographical location seems far more important than stake size (and socio-demographics; see Table 4). Further, we find a consistent pattern among Indian participants: Average offer is 0.8 tokens smaller than among US Northerners ($p = 0.001$) and, almost fittingly, responders content themselves with 0.3 tokens less ($p = 0.194$). US Southerners, on the other hand, offer 0.4 tokens less but demand 0.2 tokens more than Northerners. If one accepts Southerners' conflict as a manifestation of a regional "culture of honor" (Nisbett and Cohen, 1996), one might best interpret cross-country variation in ultimatum bargaining as a cultural difference. This finding illustrates that our online approach can be adapted to study the impact of varying cultural norms.

Table 5 demonstrates the full capability of virtual pools as we regress elicited behavior on matched regional context data. We confine our analysis to US participants. Apart from inclusion of context dummies, models are similar to Table 4.

Dictators appear strongly sensitive to both regional average income (model 1) and social capital (model 2). Regardless of operationalization preferable socio-economic conditions go along with increased generosity, adding about 8% of the endowment to average allocation. A joint analysis (model 3) suggests that both factors exert independent influence on social preferences.[15] According to our estimates, offers from wealthier and socially more integrated states are 1.3 tokens larger on average than those from less-advantaged states. Hence, total size of contextual influence clearly exceeds stake effects. Most important from a sociological perspective, context effects are both more pronounced and theoretically consistent than effects

[14] We do not claim causality for this result. Nonetheless Tan (2006) reports a similar association of religiosity and dictator behavior. Additionally, several observational studies address parents' tendencies for charitable giving (see Wiepking and Bekkers, 2012 for an overview).

[15] Variance inflation factor (VIF) is 1.25 (1.17) for the high income (social capital) dummy. If we additionally control for the American South in model 3, effects of context variables lose significance. Still, high average income ($\beta = 0.590$; $p = 0.121$) and high social capital ($\beta = 0.436$; $p = 0.314$) are more influential than the region dummy ($\beta = -0.300$; $p = 0.538$).

Table 5
Differences across context variables in the US.

	DG offer						UG-P offer		UG-R MAO	
	(1)		(2)		(3)		(4)		(5)	
	β	p	β	p	β	p	β	p	β	p
High average income	0.829	0.015*			0.689	0.045*	0.112	0.631	−0.169	0.541
High social capital			0.773	0.013*	0.608	0.049*	0.122	0.599	0.147	0.587
N	251		249		249		245		235	
R^2	0.198		0.193		0.208		0.104		0.116	

OLS regressions with sequence fixed effects and full control of socio-demographics (see Table 4). *$p < 0.05$.

of individual attributes. This is particularly true for income effects: Whereas individual disposable income is irrelevant for fairness considerations, average affluence in participants' social environment fosters giving in DG.

These findings highlight context-dependency of economic behavior in a weak experimental situation. In our stronger design (models 4 and 5), regional context ceases to influence decision-making. Still, direction of effects is consistent with theoretical expectations: Average offer is higher in wealthy and socially more integrated states (model 4); similarly, in high average income states MAO is lower (model 5). High social capital, on the other hand, appears to foster informal sanctioning of norm violations. Altogether, however, effects are small and insignificant.

Multiple shortcomings exacerbate a similar analysis for Indian participants. First, strong self-selection among workers from India (see Table 2) provides participants comparatively detached from their local environments. English-speaking university graduates (88% of our sample from India) may suffice for cross-country comparison. Given lack of representativeness, they surely do not support comparison across regions. Second, Indian states on average are more populous and more diverse than US states, providing cruder indicators of local environment. Measurement error surely weakens potential relations to elicited behavior. Third, available context data is of lower construct validity. We use state-wise gross domestic product (GDP) per capita as a proxy for average income (Government of India (2015)). Our approximation of social capital is more problematic: Lacking an analog to Putnam's (2000) index, we resorted to two complementary indicators: On the one hand, Serra (1999) provides 10 variables on state-wide political participation, trust in government officials, and membership in both religious and civic associations. On the other, Desai et al. (2010) distill an index from a national survey of more than 40,000 Indian households focusing on personal networks, organization membership, and confidence in institutions. We find no context effects of either GDP per capita or either measures of "social capital" upon experimental behavior in India (see Supplementary material for data (A7) and results (A8)). This finding does not imply the absence of regional differences—which are vast in India—and their potential effect on bargaining behavior. Instead, it mirrors the consequences of the aforementioned methodological drawbacks.

6. Discussion

Using Amazon's crowdsourcing platform *Mechanical Turk* (MTurk) we recruited a cross-national subject pool from India and the US ($N = 991$) to participate in the Ultimatum (UG) and Dictator Game (DG) at varying stake levels ($0, $1, $4, and $10). Based on a substantially more diverse sample than those obtained from traditional experimental subject pools, we were able to reproduce laboratory findings in terms of both marginal totals and treatment effects: Monetary stakes induce more rational behavior among dictators and UG responders but not for UG proposers. The overall size of stake effects is moderate, though, and rapidly saturates. Inducing $1 suffices to produce outcomes similar to (typically more strongly incentivized) laboratory set-ups. Extending prior cross-national online research (Amir et al., 2012; Raihani et al., 2013) we reproduce these findings under increased methodical equivalence: To homogenize decision situations across countries we used tokens instead of currency stakes and weighted monetary endowments in accordance with purchasing power differences (cf., Roth et al., 1991).

We further hypothesized that social norms of fairness are conditional on local context. Bringing context back into social experimentation is particularly important for sociological research. Sociologists have long been interested in the consequences of context, or "culture," for social interaction and market behavior in particular (Durkheim, 1893; Polanyi, 1944; Weber, 1920). Since then, multi-level explanation of individual action has become a disciplinary hall mark (Coleman, 1990; Merton, 1949; Parsons, 1937).

Because isolating influences of context on individual decisions is methodologically challenging empirical work lagged behind theoretical advances throughout the 20th century. Only recently have cross-country comparisons of laboratory results provided indications regarding regional variation of behavior (e.g. Brandts et al., 2004; Henrich et al., 2005; Kocher et al., 2008). Still, commentators see experimentation at multiple locations as largely speculative (e.g. Baumard and Sperber, 2010; Cardenas and Carpenter, 2008).

Online experiments thus provide a sorely needed complement (Gosling et al., 2010) to assess generalizability of empirical regularities established largely in North American and European university laboratories. Crowdsourced experimentation, however, possesses its own methodological challenges. Most worrisome is regional self-selection into participation.

Consequently, behavioral variation is only attributable to context under full control of socio-demographic composition of regional samples, or "virtual pools."

Going beyond traditional implementations of bargaining games our application revealed substantial cross-regional and cross-national variation in economic decision-making. In retrospect, one quasi-experimental finding attracts particular interest. Whereas we do not find a cross-country difference larger than regional variations in a parametric situation (DG), culture seems to be relevant in strategic interaction (UG): Participants in India are more selfish (proposers) and less demanding (responders) than US Americans. Within the US, Southerners appeared both more selfish (proposers) and more demanding (responders). This particular effect structure provides an indication of culturally heterogeneous fairness norms. As benevolence in DG does not differ between India and the American South, we reject an alternative interpretation as to which sharing of (windfall) gains might be less pronounced and little expected in poor countries. Hence, our research shows that online experimentation can be adapted to study varying cultural norms of behavior.

We were further able partly to reduce US regional variation to differences in state-wide average income and social capital. Participants from high income or high social capital contexts showed considerable stronger social preferences. Individual socio-demographics instead proved largely irrelevant. We believe that the opportunity to match context data is a key potential of crowdsourced online designs. Acknowledging local embeddedness in social experiments helps to mitigate concerns about generalizability, particularly if small-scale designs are to answer "big" questions (Jackson and Cox, 2013; Walker and Cohen, 1985). Still, the use of virtual pools—and thus the opportunity to study systematically the characteristics people bring to the experimental situation—has received scant attention in experimental research. We know a great deal about how institutional arrangements affect fairness, trust, cooperation, and reciprocity in economic games (e.g. Diekmann and Przepiorka, 2015; Fehr and Gächter, 2000; Raub and Keren, 1993), yet we know little about how local socio-economic conditions and strategies learned in daily interaction influence outcomes of social experiments.

Naturally, quasi-experiments do not provide for strict causal inference (Shadish et al., 2001), as they fail to "satisfy the strict evidential standards that the [experiment] has been set up to satisfy and, if the investigation is constrained to satisfy those standards, no ex post speculation is permitted" (Deaton, 2010, p. 441). Every sub-group analysis thus comes under suspicion of "fishing" (Shadish et al., 2001), i.e. cherry-picking data from an active search for significant differences across quasi-experimental conditions. We thus turned to two broadly reflected indicators, economic prosperity and social capital, for which theoretical expectations as to their consequences on economic behavior are rather well established. To secure internal validity, theoretical underpinning is essential in choosing context variables in applications with virtual pools. Also, analyses of virtual pools might merit reproach for reporting only cross-level correlates. Apparently, both ecological fallacy and unobserved heterogeneity potentially threaten internal validity of social aggregates' estimated "effects." Hence, we take this opportunity to formulate three desiderata for further research:

First, we argue for refined localization of participants at the level of counties, ZIP codes, or census tracts. Precise localization permits matching of lower-level context data and thus accounts for local variation of social environments. Lower aggregation allows for metrical representation of context variables and thus investigation into functional forms and trade-offs between context variables. Apart from reliance on self-administered geographical data prior studies commended storage of IP addresses for tracking and localization (e.g. Mason and Suri, 2011; Rand, 2012). Their use, however, raises ethical considerations (e.g. the necessity of participants' informed consent) and technical challenges. IP addresses (if not disguised by IP proxies) indicate Internet providers' active local network nodes at a given point in time, although not participants' exact locations.

Second, our analysis contrasted elicited behavior along differences in individual income and regional average income. Apparently, regional affluence proved more important to fairness considerations than personal income. We call for a similar comparison of effects from individual and collective social capital. Future studies should thus use extended questionnaires including Putnam's (2000) survey items and, for example, network generators quantifying participants' access to local social capital (e.g. Lin et al., 2001; Van der Gaag and Snijders, 2005).

Third, indications as to the mechanisms behind behavioral differences across regions and cultures require studies of participants in considerably more countries. Scrutinizing multiple countries is essential to isolating the effects of culture, local institutions, and economic development. Then again, inclusion of more countries occurs given selective MTurk participation, opening up a trade-off between the number of countries included and the degree of self-selection one accepts. Recent developments at Amazon confining new MTurk registration to US residents strengthen these concerns. To minimize cross-level fallacies future applications should invest in reduction of (region-specific) self-selection. A promising strategy might include stratified sampling using MTurk's filtering system, which requires participants to possess certain "qualifications" to accept a HIT.

Our study of macro-micro interrelations surely remains exploratory. Given the research gap regarding context-dependency of human behavior, our exemplary analysis of virtual pools demonstrates that crowdsourced online experiments are a promising approach to expanding the sociological experimenters' toolkit.

Acknowledgments

We thank the editor and two anonymous reviewers. Josef Brüderl and Ryan Calder kindly supported the project. We are grateful to Fabian Thiel for excellent research assistance.

Appendix A. Supplementary material

Supplementary material related to this article can be found at http://dx.doi.org/10.1016/j.ssresearch.2016.04.014.

References

Akerlof, G.A., Kranton, R.E., 2000. Economics and identity. Quart. J. Econ. 115, 715–753.
Amir, O., Rand, D.G., Gal, Y.K., 2012. Economic games on the internet: the effect of $1 stakes. PLoS One 7, e31461.
Andersen, B.S., Ertac, S., Gneezy, U., Hoffman, M., List, J.A., 2011. Stakes matter in ultimatum games. Am. Econ. Rev. 101, 3427–3439.
Andreoni, J., 1990. Impure altruism and donations to public goods: a theory of warm-glow giving. Econ. J. 100, 464–477.
Bardhan, P., Udry, C., 1999. Development Microeconomics. Oxford University Press, Oxford.
Baumard, N., Sperber, D., 2010. Weird people, yes, but also weird experiments. Behav. Brain Sci. 33, 24–25.
Berinsky, A.J., Huber, G.A., Lenz, G.S., 2012. Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. Polit. Anal. 20, 351–368.
Bicchieri, C., 2006. The Grammar of Society: The Nature and Dynamics of Social Norms. Cambridge University Press, New York.
Bicchieri, C., Xiao, E., 2009. Do the right thing: but only if others do so. J. Behav. Decis. Mak. 22, 191–208.
Blalock, H.M., 1984. Contextual-effects models: theoretical and methodological issues. Annu. Rev. Sociol. 10, 353–372.
Blau, P.M., 1960. Structural effects. Am. Sociol. Rev. 25, 178–193.
Bolle, F., 1990. High reward experiments without high expenditure for the experimenter? J. Econ. Psychol. 11, 157–167.
Bonikowski, B., 2010. Cross-national interaction and cultural similarity: a relational analysis. Int. J. Comp. Sociol. 51, 315–348.
Bourdieu, P., 1984. Distinction: A Social Critique of the Judgment of Taste. Harvard University Press, Cambridge.
Brandts, J., Saijo, T., Schram, A., 2004. How universal is behavior? a four country comparison of spite and cooperation in voluntary contribution mechanisms. Public Choice 119, 381–424.
Buhrmeister, M., Kwang, T., Gosling, S.D., 2011. Amazon's Mechanical Turk: a new source of inexpensive yet high-quality, data? Perspect. Psychol. Sci. 6, 3–5.
Camerer, C.F., 2003. Behavioral Game Theory: Experiments in Strategic Interaction. Sage, New York.
Camerer, C.F., Fehr, E., 2004. Measuring social norms and preferences using experimental games: a guide for social scientists. In: Henrich, J., Boyd, R., Bowles, S., Camerer, C.F., Fehr, E., Gintis, H. (Eds.), Foundations of Human Sociality. Oxford University Press, Oxford.
Camerer, C.F., Hogarth, R.M., 1999. The effects of financial incentives in experiments: a review and capital-labor-production framework. J. Risk Uncertain. 19, 7–42.
Cameron, L., 1999. Raising the stakes in the ultimatum game: experimental evidence from Indonesia. Econ. Inq. 37, 47–59.
Cardenas, J.C., Carpenter, J., 2008. Behavioral development economics: lessons from field labs in the developing world. J. Dev. Stud. 44, 337–364.
Carpenter, J., Verhoogen, E., Burks, S., 2005. The effect of stakes in distribution experiments. Econ. Lett. 86, 393–398.
Chandler, J., Mueller, P., Paolacci, G., 2014. Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. Behav. Res. Methods 46, 112–130.
Cohen, D., Nisbett, R.E., Bowdle, B.F., Schwarz, N., 1996. Insult, aggression, and the southern culture of honor: an 'experimental ethnography.'. J. Personal. Soc. Psychol. 70, 945–960.
Coleman, J.S., 1990. Foundations of Social Theory. Belknap Press of Harvard University Press, Cambridge.
Crump, M.J.C., McDonnell, J.V., Gureckis, T.M., 2013. Evaluating Amazon's Mechanical Turk as a tool or experimental behavioral research. PLoS One 1, e57410.
Dasgupta, P., Serageldin, I. (Eds.), 2000. Social Capital: a Multifaceted Perspective. World Bank, Washington.
Deaton, A., 2010. Instruments, randomization, and learning about development. J. Econ. Lit. 48, 424–455.
Desai, S.B., Dubey, A., Joshi, B.L., Sen, M., Shariff, A., Vanneman, R., 2010. Human Development in India: Challenges for a Society in Transition. Oxford University Press, New Delhi.
Diekmann, A., 2004. The power of reciprocity: fairness, reciprocity, and stakes in variants of the dictator game. J. Confl. Resolut. 48, 487–505.
Diekmann, A., Przepiorka, W., 2015. Punitive preferences, monetary incentives and tacit coordination in the punishment of defectors promote cooperation in humans. Sci. Rep. 5, 10321.
Duflo, E., 2007. Field experiments in development economics. In: Blundell, R., Newey, W., Persson, T. (Eds.), Advances in Economics and Econometrics. Cambridge University Press, Cambridge.
Durkheim, E., 1893. De la division du travail social. Les Presses Universitaires de France, Paris.
Elster, J., 2007. Explaining Social Behavior: More Nuts and Bolts for the Social Sciences. Cambridge University Press, Cambridge.
Engel, C., 2011. Dictator games: a meta study. Exp. Econ. 14, 583–610.
Experimental Turk, 2015. A Blog on Social Science Experiments on Amazon Mechanical Turk. https://experimentalturk.wordpress.com.
Fehr, E., Fischbacher, U., 2004. Social norms and human cooperation. Trends Cognit. Sci. 8, 185–190.
Fehr, E., Gächter, S., 2000. Cooperation and punishment in public goods experiments. Am. Econ. Rev. 90, 980–994.
Fehr, E., Schmidt, K., 1999. A theory of fairness, competition, and cooperation. Quart. J. Econ. 114, 817–868.
Forsythe, R., Horowitz, J.L., Savin, N.E., Sefton, M., 1994. Fairness in simple bargaining experiments. Games Econ. Behav. 6, 347–369.
Franzen, A., Pointner, S., 2012. Anonymity in the dictator game revisited. J. Econ. Behav. Organ. 81, 74–81.
Fukuyama, F., 1995. Trust: the Social Virtues and the Creation of Prosperity. Free Press, New York.
Goodman, J.K., Cryder, C.E., Cheema, A., 2012. Data collection in a flat world: the strengths and weaknesses of Mechanical Turk samples. J. Behav. Decis. Mak. 26, 213–224.
Gosling, S.D., Sandy, C.J., John, O.P., Potter, J., 2010. Wired but not WEIRD: the promise of the Internet in reaching more diverse samples. Behav. Brain Sci. 33, 34–35.
Government of India, 2015. Economic Survey. http://indiabudget.nic.in.
Granovetter, M., 1973. The strength of weak ties. Am. J. Sociol. 78, 1360–1380.
Gupta, N., Crabtree, A., Rodden, T., Martin, D., O'Neill, J., 2014. Understanding Indian crowdworkers. In: Proceedings of the 17th Conference on Computer Supported Cooperative Work.
Güth, W., Schmittberger, R., Schwarze, S.B., 1982. An experimental analysis of ultimatum bargaining. J. Econ. Behav. Organ. 3, 367–388.
Harkness, J.A., van de Vijver, F.J.R., Mohler, P.P. (Eds.), 2003. Cross-cultural Survey Methods. Wiley, Hoboken.
Hedström, P., Swedberg, R. (Eds.), 1998. Social Mechanisms: an Analytical Approach to Social Theory. Cambridge University Press, Cambridge.
Henrich, J., 2000. Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga. Am. Econ. Rev. 90, 973–979.
Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Hill, K., Gil-White, F., Gurven, M., Marlowe, F., Patton, J.Q., Smith, N., Tracer, D., 2005. 'Economic man' in cross-cultural perspective: behavioral experiments in 15 Small-scale societies. Behav. Brain Sci. 28, 795–855.
Henrich, J., Heine, S.J., Norenzayan, A., 2010. The weirdest people in the world? Behav. Brain Sci. 33, 61–135.
Hergueux, J., Jacquemet, N., 2015. Social preferences in the online laboratory: a randomized experiment. Exp. Econ. 18, 251–283.
Heyman, J., Ariely, D., 2004. Effort for payment: a tale of two markets. Psychol. Sci. 15, 787–793.
Hoffman, E., McCabe, K.A., Smith, V.L., 1996. On expectations and the monetary stakes in ultimatum games. Int. J. Game Theory 25, 289–301.
Horton, J.J., Rand, D.G., Zeckhauser, R.J., 2011. The online laboratory: conducting experiments in a real labor market. Exp. Econ. 14, 399–425.
Huntington, S.P., 1993. The clash of civilizations? Foreign Aff. 72, 22–49.

Inglehart, R., Baker, W.E., 2000. Modernization, cultural change, and the persistence of traditional values. Am. Sociol. Rev. 65, 19–51.
Jackson, M., Cox, D.R., 2013. The principles of experimental design and their application in sociology. Annu. Rev. Sociol. 39, 27–49.
Kocher, M.G., Cherry, T., Kroll, S., Netzer, R.J., Sutter, M., 2008. Conditional cooperation on three continents. Econ. Lett. 101, 175–178.
Lee, M.R., Bankston, W.B., Hayes, T.-C., Thomas, S.A., 2007. Revisiting the southern culture of violence. Sociol. Quart. 48, 253–275.
Levitt, S.D., List, J.A., 2007. What do laboratory experiments measuring social preferences reveal about the real world? J. Econ. Perspect. 21, 153–174.
Lin, N., 2001. Social Capital: a Theory of Social Structure and Action. Cambridge University Press, Cambridge.
Lin, N., Yang-Chih, F., Ray-May, H., 2001. The position generator: measurement techniques for investigations of social capital. In: Lin, N., Cook, K., Burt, R. (Eds.), Social Capital: Theory and Research. De Gruyter, New York.
Lynn, P., 2003. Developing quality standards for cross-national survey research: five approaches. Int. J. Soc. Res. Methodol. 6, 323–336.
Marder, J., 2015. The internet's hidden science factory. In: PBS NewsHour February 11.
Marsden, P.V., Reed, J.S., Kennedy, M.D., Stinson, K.M., 1982. American regional cultures and differences in leisure time activities. Soc. Forces 60, 1023–1049.
Mason, W., Suri, S., 2011. Conducting behavioral research on Amazon's Mechanical Turk. Behav. Res. Methods 44, 1–23.
Merton, R.K., 1949. Social Theory and Social Structure. Free Press, New York.
Morgan, S.L., Winship, C., 2007. Counterfactuals and Causal Inference: Methods and Principles for Social Research. Cambridge University Press, Cambridge.
Munier, B., Zaharia, C., 2002. High stakes and acceptance behavior in ultimatum bargaining. Theor. Decis. 53, 187–207.
Nisbett, R.E., Cohen, D., 1996. Culture of Honor: The Psychology of Violence in the South. Westview Press, Boulder.
OECD, 2013. Country Statistical Profile: India. http://dx.doi.org/10.1787/csp-ind-table-2013-2-en.
Oosterbeek, H., Sloof, R., van de Kuilen, G., 2004. Cultural differences in ultimatum game experiments: evidence from a meta-analysis. Exp. Econ. 7, 171–188.
Ostrom, E., Ahn, T.K. (Eds.), 2003. Foundations of Social Capital. Edward Elgar, Cheltenham.
Paolacci, G., Chandler, J., 2014. Inside the turk: understanding Mechanical Turk as a participant pool. Curr. Dir. Psychol. Sci. 23, 184–188.
Parsons, T., 1937. The Structure of Social Action. Macmillan, New York.
Polanyi, K., 1944. The Great Transformation. Farras & Rinehart, New York.
Putnam, R.D., 1993. Making Democracy Work: Civic Traditions in Modern Italy. Princeton Universitiy Press, Princeton.
Putnam, R.D., 2000. Bowling Alone: the Collapse and Revival of American Community. Simon & Schuster, New York.
Raihani, N.J., Mace, R., Lamba, S., 2013. The effect of $1, $5 and $10 in an online dictator game. PLoS One 8, e73131.
Rand, D.G., 2012. The promise of Mechanical Turk: how online labor markets can help theorists run behavioral experiments. J. Theor. Biol. 299, 172–179.
Rand, D.G., Peysakhovich, A., Kraft-Todd, G.T., Newman, G.E., Wurzbacher, O., Nowak, M.A., Greene, J.D., 2014. Social heuristics shape intuitive cooperation. Nat. Commun. 5, 3677.
Raub, W., Keren, G., 1993. Hostages as a commitment device: a game-theoretic model and an empirical test of some scenarios. J. Econ. Behav. Organ. 21, 43–67.
Rauhut, H., Winter, F., 2010. A sociological perspective measuring social norms by means of strategy method experiments. Soc. Sci. Res. 39, 1181–1194.
Reips, U.-D., 2002. Standards for internet-based experimenting. Exp. Psychol. 49, 243–256.
Rhodes, S.D., Bowie, D.D., Hergenrather, K.C., 2003. Collecting behavioral data using the world wide web: considerations for researchers. J. Epidemiol. Community Health 57, 68–73.
Robinson, W.S., 1950. Ecological correlations and the behavior of individuals. Am. Sociol. Rev. 15, 351–357.
Ross, J., Irani, I., Silberman, M.S., Zaldivar, A., Tomlinson, B., 2010. Who are the crowdworkers? shifting demographics in Amazon Mechanical Turk. Proc. AMC CHI Conf. 2010, 2863–2872.
Roth, A., Prasnikar, V., Okuno-Fujiwara, M., Zamir, S., 1991. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh and Tokyo: an experimental study. Am. Econ. Rev. 81, 1068–1095.
Sampson, R.J., Morenoff, J.D., Gannon-Rowley, T., 2002. Assessing 'neighborhood effects': social processes and new directions in research. Annu. Rev. Sociol. 28, 443–478.
Serra, R., 1999. 'Putnam in India': Is Social Capital a Meaningful and Measurable Concept at Indian State Level? Working Paper 92 Institute of Development Studies, University of Sussex.
Shadish, W.R., Cook, T.D., Campbell, D.T., 2001. Experimental and Quasi-experimental Designs for Generalized Causal Inference. Houghton Mifflin, Boston.
Slonim, R., Roth, A.E., 1998. Learning in high stakes ultimatum games: an experiment in the Slovak Republic. Econometrica 66, 569–596.
Tan, J.H.W., 2006. Religion and social preferences: an experimental study. Econ. Lett. 96, 133–139.
Tocqueville, A., 1840. Democracy in America. Saunders and Otley, London.
US Bureau of Economic Analysis, 2015. Regional Data: GDP and Personal Income. http://bea.gov/iTable/index_regional.cfm.
Van der Gaag, M., Snijders, T., 2005. Resource generator: measurement of individual social capital with concrete items. Soc. Netw. 27, 1–29.
Walker, H.A., Cohen, B.P., 1985. Scope statements: imperatives for evaluating theory. Am. Sociol. Rev. 50, 288–301.
Weber, M., 1920. Die protestantische Ethik und der Geist des Kapitalismus. J.C.B. Mohr, Tübingen.
Weinberg, J.D., Freese, J., McElhattan, D., 2014. Comparing data characteristics and results of an online factorial survey between a population-based and a crowdsource-recruited sample. Sociol. Sci. 1, 292–310.
Wells, J.S., Rand, D.G., 2013. Strategic Self-interest Can Explain Seemingly 'fair' Offers in the Ultimatum Game (SSRN Working Paper).
Wiepking, P., Bekkers, R., 2012. Who gives? A literature review of predictors of charitable giving. Part two: gender, family composition and income. Volunt. Sect. Rev. 3, 217–245.
Williamson, V., 2014. On the Ethics of Crowdsourced Research. Working Paper. Harvard.