On the Transportability of Laboratory Results

Felix Bader, Bastian Baumeister, Roger Berger and Marc Keuschnigg

The self-archived postprint version of this journal article is available at Linköping University Institutional Repository (DiVA):

<http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-153265>

N.B.: When citing this work, cite the original publication.

Bader, F., Baumeister, B., Berger, R., Keuschnigg, M., (2019), On the Transportability of Laboratory Results, *Sociological Methods & Research*.

Original publication available at:

Copyright: SAGE Publications (UK and US)

<http://www.uk.sagepub.com/home.nav>

On the Transportability of Laboratory Results

Felix Bader¹, Bastian Baumeister², Roger Berger², and Marc Keuschnigg^{3,*}

¹School of Social Sciences, University of Mannheim, A5 6, 68159 Mannheim, Germany

²Institute of Sociology, University of Leipzig, Beethovenstrasse 15, 04107 Leipzig, Germany

³Institute for Analytical Sociology, Linköping University, Norra Grytgatan 10, 601 74 Norrköping, Sweden. * Corresponding author; marc.keuschnigg@liu.se

Forthcoming in *Sociological Methods & Research*

Abstract

The “transportability” of laboratory findings to other instances than the original implementation entails the robustness of rates of observed behaviors and estimated treatment effects to changes in the specific research setting and in the sample under study. In four studies based on incentivized games of fairness, trust, and reciprocity, we evaluate (1) the sensitivity of laboratory results to locally recruited student-subject pools, (2) the comparability of behavioral data collected online and, under varying anonymity conditions, in the laboratory, (3) the generalizability of student-based results to the broader population, and (4), with a replication at Amazon Mechanical Turk, the stability of laboratory results across research contexts. For the class of laboratory designs using interactive games as measurement instruments of prosocial behavior we find that rates of behavior and the exact behavioral differences between decision situations do not transport beyond specific implementations. Most clearly, data obtained from standard participant pools differ significantly from those from the broader population. This undermines the use of empirically motivated laboratory studies to establish descriptive parameters of human behavior. Directions of the behavioral differences between games, in contrast, are remarkably robust to changes in samples and settings. Moreover, we find no evidence for either anonymity effects nor mode effects potentially biasing laboratory measurement. These results underscore the capacity of laboratory experiments to establish generalizable causal effects in theory-driven designs.

Keywords

Anonymity, experimental methods, external validity, laboratory research, mode effects, online experiments, prosocial behavior, sample effects

Acknowledgments

We thank Peter Hedström, Karl-Dieter Opp, Merlin Schaeffer, Tobias Wolbring, and three anonymous reviewers for valuable comments. We are grateful to Hanna Nau, Leona Przechowski, Lennart Rösemeier, Fabian Thiel, Janine Thiel, and Anna Wolf for excellent research assistance and to Marion Apelt and Regina Heindl for administrative support. This project received financing through generous grants from the German Research Foundation (BE 2372/3-1 and KE 2020/2-1).

M.K. further acknowledges funding from the European Research Council (324233), the Swedish Research Council (445-2013-7681, 340-2013-5460), and Riksbankens Jubileumsfond (M12-0301:1). F.B. and M.K. contributed equally to this work.

1 Introduction

Laboratory experiments have a decisive methodological advantage over alternative modes of data generation in the social sciences: Group formation, randomization, and manipulation—while holding environmental factors constant—ease the testing of hypotheses regarding causes and effects (e.g., Falk and Heckman 2009; Shadish, Cook, and Campbell 2002; Webster and Sell 2014). Because of their support for causal inference (internal validity), many consider laboratory experiments the “gold standard” of scientific inquiry (e.g., Morgan and Winship 2015; Rubin 2008). In theory, these benefits of internal validity also hold for randomized field experiments (Gerber and Green 2012) and carefully chosen natural experiments (Dunning 2012). Due to its potential of absorbing confounders, however, benefits are most pronounced for experiments implemented in the artificial environment of a laboratory. Note that our discussion of social experiments focuses on designs measuring actual behavior rather than behavioral intentions, attitudes, or opinions. In this tradition, laboratory research allows the elimination of plausible alternative explanations for results and generalization is directed to the support or nonsupport of theoretical principles (Thye 2014; Willer and Walker 2007; Zelditch 2014).

Others have criticized laboratory research in the social sciences due to its often questionable generalizability to the “real world” (external validity). In social science lab research, external validity refers first and foremost to lab–field generalizability and thus to the question whether individuals examined in the laboratory behave as they would in everyday life (Jackson and Cox 2013; Levitt and List 2007). This question is particularly relevant if one conceives of laboratory methods also as measuring instruments of certain types of behavior (e.g., Franzen and Pointner 2013; Glaeser et al. 2000; Rauhut and Winter 2010). Upstream requirements for external validity entail that laboratory results are robust to changes in both the specific research setup and the sample under study (Campbell and Stanley 1963; Cronbach 1982). These criteria convey the “transportability” (Pearl and Bareinboim 2014) of findings to other implementations beyond any specific design. After all, “[m]ost experiments are highly local but have general aspirations” (Shadish et al. 2002:18).

This article tests the minimal requirements for external validity of laboratory research in the social sciences. We conceptualize a laboratory design as a specific combination of subjects (*units*), stimuli (*treatments*), measurements (*observations*), and context (*setting*). This decomposition was first introduced by Cronbach (1982:78) and has been used by others (e.g., Gerber and Green 2012; Shadish et al. 2002) to evaluate experimental findings’ range of validity.

In four studies, we assess each dimension’s importance for establishing transportability: Study 1 varies units in a multi-location laboratory comparison conducted at two German

universities, in Leipzig and Munich (*pool generalizability*). Study 2 targets observations and tests for comparability of behavioral data collected online and, under varying anonymity conditions, in the laboratory (*mode generalizability*). Study 3, a nationwide online implementation, again relates to units and tests our baseline results’ transportability to the broader population (*sample generalizability*). Study 4 concerns the setting of data collection (*context generalizability*) and—transporting our standardized decision situation into an online labor market—considers workers at Amazon’s crowdsourcing platform Mechanical Turk (MTurk). Different samples, modes, or settings may violate transportability in that they produce different rates of observed behaviors and—more worryingly for experimental research—heterogeneous treatment effects.

Our results ground on behavioral data collected in incentivized games of fairness, trust, and reciprocity from 2,664 subjects using the same decision interface. Throughout our four studies, we focus on two decision-making situations frequently used in the methodological research on laboratory designs: the Dictator Game (DG) and the Trust Game (TG). These games differ in complexity, carry the potential for socially desirable responses, and enable a direct comparison with an extant literature. Because socially acceptable (DG) and socially optimal (TG) behaviors diverge from first movers’ egoistic strategies, these games reveal expectations about valid norms of fairness (DG), trust, and reciprocity (TG) in a particular population and setting (Bicchieri 2006; Elster 2007). We compare behavior in these situations replicating the common finding that an investment opportunity (TG), rather than altruism (DG), motivates first movers to higher transfers. Our focus on the interplay between games advances prior studies on the transportability of lab results, allowing us to investigate how qualitative results (the ranking of mean transfers across games) and point estimates of behavioral differences between decision situations (the within-subject differences in transfers across games) generalize to other units, observations, and settings.

Detecting violations of transportability in laboratory designs constitutes a lively research area within experimental economics. This activity has led to significant advances in the way social scientists implement and interpret laboratory studies (see the overviews by Fréchette 2016; Galizzi and Navarro-Martínez 2018; Levitt and List 2007). These efforts have, however, remained selective, focusing on particular aspects of laboratory designs one at a time, such that results regarding lab findings’ sensitivity toward changes in implementation are often mixed. The “replication crisis” in the social sciences (Chang and Li 2015; Freese 2007; Open Science Collaboration 2015) reinforces the call for thorough tests of experimental reliability, more diverse samples, and taking into account of potentially heterogeneous treatment effects. The narrow variation of socio-demographics in standard experimental subject pools remains conspicuous (Druckman and Kam 2011; Henrich, Heine, and Norenzayan 2010; Peterson 2001), particularly when experimenters seek general insights into human behavior or estimate treatment effects that may interact with individuals’ background characteristics.

Against this backdrop, we systematically assess the transportability of laboratory results and map safe grounds for behavioral studies conducted both in the laboratory and online. In the remainder, we proceed as follows: In section 2, we use Cronbach’s (1982) decomposition of laboratory designs into units, treatments, observations, and settings to delineate how each dimension relates to the general desideratum of transportability. In section 3, we discuss

established protocols useful in identifying threats to external validity in social science lab research. This review will motivate our own test strategies and highlight how our four studies complement the existing literature. In section 4, we outline our design. Section 5 presents our results. In the concluding section, we discuss our findings’ practical implications for experimenters in the social sciences.

2 Demands to Transportability

Cronbach (1982) defines laboratory designs as combinations of specific units, treatments, observations, and settings. The acronym *utos* refers to the particular “instances on which data are collected” (p. 78). Each dimension has consequences for the transportability of laboratory results (see also Shadish et al. 2002; we follow their simplified conceptualization).

Units refer to the participants of a laboratory study. For external validity, participants must be broadly representative of the target population to which one wishes to generalize. This implies random sampling from the target population and the use of inference statistics. Generalizability under non-random sampling requires—as a minimal condition—that socio-demographic characteristics relevant to sustaining expected treatment effects overlap in the subject pool and the target population.

Treatments represent the randomized stimuli participants are exposed to. Treatments should reproduce real-world conditions as closely as possible. It is most important for theory-driven experiments, however, that treatments closely represent the theoretical concepts under study (construct validity). In addition, treatments should be well calibrated. Subtle treatments, for example, induce the risk of experimenters mistaking a lack of treatment perception for a null result (treatment validity).

Observations denote the measurement of outcome variables. In laboratory studies, reactivity is a major measurement concern. Subjects’ feeling of being observed can shift measurements toward socially desirable outcomes (Pygmalion effect), and subjects potentially bias measurements by forming beliefs about the purpose of scientific inquiry (experimenter demand effect). Online experiments, which have recently become popular, offer increased anonymity but less control over the participants’ surroundings.

Settings characterize the context of data generation. Transportability, again, relates to the mapping of the laboratory setup to real-world conditions. The artificial lab context generally runs counter to this criterion but at least allows for “experimental realism:” Experimenters must implement theoretically relevant features in a way that allows participants to assign similar meanings as they would in natural contexts.

Cronbach’s dimensions indicate the range of validity for a given laboratory study. Results directly generalize to units, treatments, observations, and settings fully covered in the experiment (*utos*). Generalizations to conditions beyond those covered in the laboratory

(which Cronbach terms *UTOS*) require additional bridging assumptions. The range of validity straddles, however, for conditions clearly deviating from the empirical implementation in at least one of these four dimensions. In such cases, Cronbach speaks of **UTOS*. Here, the laboratory design no longer sustains transportability.

The main purpose of experimental research is to test causal relationships derived from theoretical hypotheses (Martin and Sell 1979; Willer and Walker 2007). External validity, however, is compromised if elements not randomized by the design, such as population, period, or setting, interact with the hypothesis under study (Zelditch 2014). Ideally, theory should inform about the scope of its application and delineate potential heterogeneous treatment effects to enable valid experimental tests. If underlying theories are incomplete—as in the case of “effect experiments” (Zelditch 2014:183)—the challenge of establishing external validity is much greater (Schram 2005). Behavioral economists, for example, frequently measure rates of behavior in incentivized decision situations from convenience samples in order to generalize regularities of human behavior (e.g., “social preferences”) or infer effects of “culture” on observed rates of behavior (see Kessler and Vesterlund 2013; Levitt and List 2007 for critique). Among other things, our research will underscore the problems associated with such empirically driven applications of laboratory designs.

3 Prior Results and Our Contribution

Following Cronbach’s (1982) typology as a structuring framework, we briefly discuss established designs for identifying threats to the transportability of laboratory results. We mainly draw on studies from experimental economics which, in the last decade, saw lively research on the methodological issues of laboratory research. We restrict our review to studies identifying potential violations based on protocols measuring prosocial behavior¹ and highlight how our four studies advance prior work.

3.1 Units

Multi-location experiments evaluate the sensitivity of laboratory results to locally recruited subject pools. Roth and colleagues’ (1991) parallel implementation of bargaining games at universities in Jerusalem, Ljubljana, Pittsburgh, and Tokyo is a classic in this domain. Close monitoring of local experimenters, careful translations, and the adjustment of stakes according to differences in purchasing power led the authors to conclude that “[b]ecause of the way the experiment was designed [...] the differences in bargaining behavior among countries are not due to differences in languages, currencies, or experiments but may tentatively be attributed to cultural differences” (p. 1068). Many studies followed (e.g., Brandts, Saijo, and Schram 2004; Henrich et al. 2001; Kocher et al. 2008), comparing elicited behaviors across locations; yet—just as in Roth et al. (1991)—they confound local pool effects with differences in nationality and culture. We fill this gap in Study 1, comparing laboratory results across student-subject pools at two German universities in Leipzig and Munich.

A second design evaluates whether lab results from student participants transport to broader, more representative populations. Studies of this type invite non-student residents

from the proximity of a university to participate in lab sessions (Anderson et al. 2013; Belot, Duch, and Miller 2015; Cappelen et al. 2015; Falk, Meier, and Zehnder 2013) and compare the results to control sessions featuring student participants.² These comparisons find students less generous, trustful, and cooperative than their non-student counterparts. Apparently, social-preference parameters estimated from student pools do not generalize to more general samples.³ We complement these efforts in Study 3, comparing our student baseline to findings from a nationwide implementation conducted over the Internet.

3.2 Observations

Subjects' feelings of being observed can shift measurements toward socially desirable outcomes. A common strategy to assess reactivity in the laboratory relies on the variation of anonymity conditions. Extending from standard setups—which protect anonymity toward other subjects—Franzen and Pointner (2012) and Hoffman, McCabe, and Smith (1996) use procedures which ensure anonymity toward the experimenter as well (using blinds, anonymized envelopes, or randomized-response techniques). Both studies report decreased rates of socially desirable behavior under increased anonymity. Barmettler, Fehr, and Zehnder (2012), on the other hand, find no effect of anonymity toward the experimenter. We address subject reactivity with a manipulation of anonymity conditions in our two laboratories. If subjects' feeling of being observed affects measurements in the laboratory, we should encounter less prosocial behavior with rising anonymity levels. Our manipulation does not aim at testing theoretical explanations of prosocial behavior (e.g., social control vs. internalized norms) but tests the comparability of data generated under different anonymization procedures commonly used in social science lab research.

Anonymity is also attainable through online experiments. These have become increasingly popular among social scientists due to both low costs and access to broad participant pools (e.g., Gosling et al. 2010; Rand 2012). Online experimenters, however, obtain no direct control over the participants' surroundings, which may pose threats to internal validity (Clifford and Jerit 2014; Reips 2002). For example, subjects may find themselves observed by others during participation, search the Internet for eligible strategies, or disbelieve the supposed interaction with other human subjects. A rigorous test strategy for mode effects of data collection requires members of the same population to take part in the same study in either the lab or the online version. Drawing on student participants, Beramendi, Duch, and Matsuo (2016) find no mode effect on various outcome measures, including the DG and a modified version of the Public Goods Game. The authors, however, failed to randomize subjects effectively, leading to marked socio-demographic differences between lab and online participants. Hergueux and Jacquemet (2015), on the other hand, randomized students to parallel lab or online sessions. Their study finds higher rates of selfish behavior among lab subjects. In their study, however, online participants received payoff through PayPal and—being spared traveling to the physical lab—faced lower participation effort. We fix these issues in Study 2, randomizing student subjects into either lab or online sessions while keeping participation effort constant across modes. We then compare our online results to lab results obtained under varying anonymity conditions.

3.3 Settings

Manipulations of the setting of data generation address the crucial issue of “real-world” generalizability. A growing number of studies compares lab behavior with choices made in concealed field experiments. In a rigorous variant of this design, researchers take efforts closely to map the artificial decision space (e.g., DG) onto the unobtrusive measurement (e.g., giving to a charity) and then exploit within-subject comparisons between settings. Typically, these studies find qualitative lab–field correspondence: Individuals who share, cooperate, or trust in the lab also exhibit more prosocial behavior in the field (e.g., Benz and Meier 2008; Englmaier and Gebhardt 2016; Franzen and Pointner 2013). Some implementations, however, report zero correlations (for a review see Galizzi and Navarro-Martínez 2018) and, more importantly, the empirical evidence at hand is likely to suffer from publication bias (Coppock and Green 2015). An alternative design addressing the realism of experiments utilizes the sampling of professionals with relevant task experience (e.g., Alevy, Haigh, and List 2007; Fehr and List 2004; Potters and van Winden 2000) in “framed field experiments” (Harrison and List 2004:1014): Because legislators, managers, and traders import their day-to-day experiences into the experimental situation, instructions can trigger work-related frames and heuristics altering the context of the experiment (Fréchette 2015).

A related and recently popularized strategy to vary experimental settings makes use of the large and heterogeneous participant pool sustained at MTurk (e.g., Amir, Rand, and Gal 2012; Berinsky, Huber, and Lenz 2012; Crump, McDonnell, and Gureckis 2013). Many consider the platform a real online labor market (Horton, Rand, and Zeckhauser 2011; Rand 2012) in which workers seek profit-maximizing allocation of time and qualification. In addition, many workers at MTurk are experienced participants in social experiments (Chandler, Mueller, and Paolacci 2014; Rand et al. 2014) and the perceived social distance is likely to be larger among MTurk participants than among traditional laboratory subjects. As a result, experimenters can expect to observe different and more “rational” situational logics than what one is used to from physical laboratories. In Study 4, we replicate our online implementation at MTurk to test the robustness of behavioral data collection against a change in the research setting.

3.4 Treatments

Methodological research on laboratory designs makes frequent use of two decision situations, the Dictator Game (DG) and Trust Game (TG). In each situation, participants must choose between self-interested and socially desirable behaviors. The games are thus natural candidates for investigations into anonymity effects, mode effects, and the sensitivity of results to different samples and contexts.

In DG, a participant receives a monetary stake and can decide how much of the pie (0–100%) she passes to a receiver (Kahneman, Knetsch, and Thaler 1986). Experimenters typically interpret giving as a manifestation of prosocial preferences. TG, on the other hand, mimics an investment decision, thereby introducing the possibility of non-reciprocity by a second mover (Berg, Dickhaut, and McCabe 1995): A trustor and a trustee each receive a stake. The trustor can decide how much of her stake (0–100%) she sends to the trustee.

The experimenter doubles this amount. The trustee then decides how much of the doubled amount (0–100%) she sends back to the trustor. Placing trust depends on the trustor’s belief in the validity of a prosocial norm of reciprocity securing trustee’s trustworthiness. Unlike in DG, first movers are required to form expectations on second movers’ likelihood of reciprocation (Glaeser et al. 2000).

The two decision situations do not qualify as experiments due to their lack of treatments. The interplay between games, however, allows us to replicate the common finding (e.g., Camerer 2003; Camerer and Fehr 2004) that an investment opportunity (TG) motivates first movers to higher transfers than altruism (DG). We expect first movers to share more in TG than in DG. Specifically, we test whether the differences in mean transfers transport to different samples, modes, and settings. Substituting TG for DG varies a bundle of aspects (e.g., parametric vs. interactive decision situation, endowment for one vs. two players) and, hence, our variation does not permit isolation of a narrow causal effect that is more typical of the sociological literature using experiments. Still, the within-subject comparison across games provides an estimate of the “treatment” effect of changing from one decision situation to another. Our focus on this interplay extends prior studies of lab results’ transportability, as it allows us to investigate how qualitative results (the ranking of transfers across games) and behavioral differences between laboratory conditions (the within-subject differences in transfers across games) generalize to other units, observations, and settings.

4 Design

Table 1 summarizes the different study designs (see Appendix A1 for sample descriptives). We first describe the sampling of participants. Our procedures then include randomization, instructions, incentives, collection of survey data, and payoff.

[Table 1 about here]

Sampling. For Studies 1 and 2, we established two student-subject pools at universities in Leipzig and Munich. We standardized recruiting across both locations advertising sign-up in introductory lectures, campus cafeterias, and university websites. From each pool, we randomly selected registered students to participate in a given lab or online session synchronized across locations. In Study 3, we examine a cross-section of the German population sampled from Forsa’s offline-recruited online access panel. Forsa uses county-level random digit dialing to register participants who privately use the Internet at least once a week. Our sample is representative of the German-born population with regard to gender, age, and administrative district, and highly heterogeneous with regard to education, occupation, and income. In Study 4, we replicate our setup at MTurk, recruiting workers from the United States and from India. Both countries make up the largest shares of platform participants (Ipeirotis 2017). For each country, we advertised participation twice per day (early morning and late afternoon local time).

Randomization. Each subject participated in DG and TG. We randomized participants to sequences of games and first- and second-mover roles. The absence of feedback in-between games secured independence of sequential behavior, enabling within-subject comparison of decision situations. To neutralize reputation effects, we randomly matched participants to another anonymous participant for each decision.⁴

Instructions. We standardized the decision interface in our four studies using a web-browser implementation based on the package SoSci Survey (www.soscisurvey.de). Our instructions map participants’ choices to payoffs as clearly as possible using GIF-animated examples but avoiding suggestion of specific strategies or frames. We only allowed individual transfers in each game to be multiples of 10% of the endowment (including 0%). In Studies 1–3, we used instructions in German; Study 4 uses similar instructions in English (see Appendix A6). We monitored understanding using control questions following each decision.

Incentives. In Studies 1 and 2, we incentivized DG with 10€; in TG, each player received an endowment of 5€. Rather than keeping stakes constant across samples, we chose monetary incentives typical for the respective participant pool to counter self-selection based on monetary motivations; in Studies 1 and 2 stakes also need to cover subjects’ effort to travel to the laboratory. In Study 3, DG was worth 5€ and each player in TG received an endowment of 2.5€. In Study 4, stakes were US\$2 in DG and US\$1 in TG. Critics may find fault at our heterogeneous stake levels pointing to the idea that observed prosociality may decrease in stake sizes. Prior evidence from laboratory (e.g., Camerer and Hogarth 1999; Carpenter, Verhoogen, and Burks 2005) and online studies (e.g., Amir et al. 2012; Keuschnigg, Bader, and Bracher 2016), however, indicates that—although monetary stakes increase selfishness compared to unincentivized games—differences in positive stakes have negligible effects on laboratory results in fairness and cooperation research.

Survey data. We requested each participant to fill out a questionnaire including items on socio-demographics, experimental experience, and—in our online Studies 2–4—the physical and social surroundings during participation (see Appendix A1). We administered the questionnaire at the end of each session.

Payoff. To compute individual payoff, we randomly drew one of ego’s (and partner’s) decisions in the games. We made randomized rewards (Bolle 1990) common knowledge in our instructions, explaining that each decision could fully determine a participant’s reward. In Studies 1 and 2, we paid participants in cash at the end of each session. Payoff included a fixed show-up fee of 2.50€ and additional earnings of 5.13€ on average (min=0.00, max=15.00). In Study 3, participants received payoff in the form of an Amazon voucher, complying to Forsa’s standard payment scheme. We set the show-up fee to 2.00€, additional earnings average 2.83€ (min=0.00, max=7.50). In Study 4, workers received payoffs via MTurk. As typically done in online experiments at MTurk, we chose a show-up fee of US\$1, additional earnings average US\$0.85 (min=0.00, max=3.00).

We introduced additional manipulations to identify both anonymity effects in laboratory data collection and a potential mode effect between laboratory and online data collection. We randomized each experimental treatment on the session level.

Anonymity. Each lab participant in Leipzig and Munich was presented with one of three anonymity conditions (see Appendix A7 for photographic documentation). (1) *Low anonymity*: In this control condition, workplaces had no shielding and participants could see one another while taking decisions ($N=115$ in Leipzig and 113 in Munich). After completion, the experimenter called each participant by her seat number to receive payoff individually at the experimenter’s desk. (2) *Standard anonymity*: Blinds shielded each workplace to create inter-subject anonymity ($N=116$ in Leipzig and 122 in Munich). After completion, we followed the above payoff procedure. This setup is typical for most laboratory implementations of social experiments. (3) *High anonymity*: We also placed the experimenter behind a blind to prevent visual contact throughout data collection ($N=131$ in Leipzig and 116 in Munich). After completion, the experimenter called participants by their seat numbers and each subject received payoff individually in a designated payment room outside the lab from a person who did not appear as an experimenter in the process of the experiment. This person sat behind a closed door with a mail slot through which each participant handed over her seat number and received payoff in an anonymized envelope. This setup creates anonymity toward both other participants and the experimenter. We made the respective anonymity condition common knowledge upon arrival. For treatment validity, we provided a detailed description of the relevant scheme in our opening instructions ensuring complete understanding of the setup.

Modes. We randomized student subjects in Leipzig and Munich to participate in either a laboratory or an online session. To avoid self-selection, we informed participants only after enrollment for a certain session about the respective mode of data collection. We held online sessions simultaneously to our laboratory sessions, thus neutralizing the “mode selection effect” and isolating the “mode measurement effect” (Hox, de Leeuw, and Klausch 2017:511). To homogenize participation effort, online participants ($N=122$ in Leipzig and 115 in Munich) had to collect their payoff in cash within one week after completion from the respective university’s laboratory, where we followed the high-anonymity payment scheme outlined above (about which online participants knew upon entering the experiment). Apart from identifying mode effects, this treatment permits a rigorous isolation of pool effects: Online implementation absorbs potential effects from both local experimenter characteristics and the two laboratories’ overall physical appearance.

5 Results

Figure 1 summarizes dictators’ and trustors’ average transfers (as percentages of their individual endowments) across studies. Pooled across samples, modes, and research contexts,

the mean allocation in DG is 42.2%. Changing from DG to TG increases average transfers by 10.4 percentage points to 52.6%.⁵

[Figure 1 about here]

To evaluate transportability of quantitative results, we test their sensitivity to locally recruited student-subject pools (Study 1), the comparability of behavioral data collected online and, under varying anonymity conditions, in the laboratory (Study 2), the generalizability of elicited behavior from student participants to the broader population (Study 3), and the stability of results across settings (Study 4). This entails running seven pairwise comparisons for both DG and TG as indicated in Figure 1. We report p -values of two-sided t -tests with robust standard errors throughout. To account for different socio-demographic compositions in our samples, we further adjust our measures by a list of participants' background characteristics (see Appendix A2 for model specification). To speak of cross-sample differences in rates of observed behavior, those need to resist conditioning. Five out of seven pairwise comparisons return significant differences for DG, but only two out of seven do so for TG.⁶

Pool generalizability. In Study 1, we find marked differences in mean DG allocations across student pools: In Leipzig, dictators allocate 42.7% on average; in Munich, they share only 35.5% of their endowment (see shaded bars in Figure 1, left panel). This gap remains after controlling for socio-demographic differences in local pool composition (blank bars; $t=4.44$, $p<.001$). In TG, Leipzig students transfer 53.5% on average; in Munich, this rate is 49.3%. These rates are not significantly different under conditioning on socio-demographics (blank bars in Figure 1, right panel; $t=1.76$, $p=.079$). Our synchronized test thus establishes pool generalizability for investment decisions. For altruistic donations, however, elicited behavior varies considerably between locations.

Mode generalizability. In Study 2, we find no evidence for a mode effect of data collection. For both games, elicited behavior is not significantly different irrespective of whether we run the study in a laboratory or online. This holds for student pools in both Leipzig ($t=.36$, $p=.721$ in DG; $t=1.52$, $p=.129$ in TG) and Munich ($t=.27$, $p=.785$ in DG; $t=.02$, $p=.987$ in TG). These results further substantiate the cross-location difference found for DG in Study 1: By shutting off potential experimenter effects and differences in labs' physical appearance, the online design identifies the gap between locations as a genuine pool effect. Furthermore, because the gap resists conditioning on socio-demographics, unstable results across locations obviously do not stem from different pool compositions.

[Figure 2 about here]

Anonymity. In Figure 2, we compare online results to the three anonymity conditions participants faced in our physical labs.⁷ At each location, lab results do not differ across anonymity conditions. Setups creating anonymity toward other participants (standard anonymity; $t=.48$, $p=.633$ in DG; $t=1.13$, $p=.261$ in TG) and, additionally, toward the

experimenter (high anonymity; $t=.04$, $p=.966$ in DG; $t=.61$, $p=.541$ in TG) do not yield different results than a low-anonymity setup. Similarly important, the results from either anonymity condition do not differ significantly from our online implementations at both locations.⁸ Anonymity effects, it seems, are not a major concern for laboratory research.

Sample generalizability. In Study 3, we contrast student-based results to those obtained in the broader population (Figure 1). Because we ran our nationwide study over the Internet, our online results among students provide the relevant benchmark. Even after controlling for socio-demographic differences, results for students (39.4% in DG; 48.4% in TG) do not generalize to a broader population sample, whose members on average share significantly more in both DG (47.7%; $t=3.72$, $p<.001$) and TG (58.8%; $t=3.20$, $p=.001$). Differences in comprehension of instructions may further aggravate direct comparisons between student and non-student samples. To test whether difficulties in understanding drive prosocial choices in our nationwide sample, we introduced a time-pressure/time-delay treatment for non-student participants. We report these results in Appendix A3 and find no statistically significant effect of this manipulation, suggesting that difficulties in understanding do not explain higher rates of prosocial behavior among non-students.

Context generalizability. In Study 4, we replicate our online implementation at MTurk to test the robustness of behavioral data collection against a change in the research setting (Figure 1). On average, crowdworkers allocate 33.7% in DG and transfer 42.4% in TG. Both rates are lower than the quantitative results obtained in our university-implemented setting using volunteer student participants ($t=2.41$, $p=.016$ in DG; $t=1.76$, $p=.078$ in TG) and participants from the broader population ($t=9.61$, $p<.001$ in DG; $t=8.09$, $p<.001$ in TG). Quantitative results, already heterogeneous across student and non-student samples, apparently do not transport to another setting. MTurk workers use more “rational” situational logics than we find among either students or members of the wider population in Germany. We find only little evidence, however, for different decision-making by experienced participants: Non-naïve subjects in all studies, on average, share less in DG—but only 0.005 percentage points per prior experiment—while experience has zero effect in TG (see Appendix A2). Note that our results are robust to the adjustment for experience. Hence, experience cannot explain the behavioral differences between the crowdworkers and the participants in our remaining studies.

[Figure 3 about here]

In Figure 3, we test for the stability of behavioral differences between decision situations across samples, modes, and settings. We focus on conditional means, keeping socio-demographic composition constant across studies. Shaded bars show average DG allocations. Blank bars on top represent first-mover transfers in TG. The difference between shaded and blank bars shows by how much, in each study, TG transfers exceed DG allocations. We

include 95% confidence intervals for the difference in average transfers between DG and TG. Unlike our statistical tests above, which we based on between-subject comparison, we now use variation within subjects to test for significance of this “treatment” effect. Our main qualitative result is robust to changes in units, observations, and settings: In each Study, average TG transfers exceed DG allocations substantially and significantly. The exact size of this difference, however, varies considerably. Among student participants in Studies 1 and 2, differences between DG and TG are small in Leipzig (8.6 percentage points, $t=3.78$, $p<.001$) but large in Munich (12.0 percentage points, $t=5.46$, $p<.001$). In Munich, TG substantially increases transfers, compensating for the lower propensity to share in a situation of altruism. Using mean DG allocations as a baseline, the change to TG raises average transfers by 33.1% in Munich, but only by 20.2% in Leipzig ($t=1.96$, $p=.050$). TG in the nationwide sample raises average sharing as measured in DG by 23.3% ($t=8.43$, $p<.001$) and, at MTurk, by 25.6% ($t=5.16$, $p<.001$).

6 Implications

In laboratory research, the benefits of artificiality—systematic variation of experimental conditions, control of confounders, and replicability—trade off with generalizability to the “real world.” We used Cronbach’s (1982) decomposition of experiments into units, treatments, observations, and settings to identify those parts of laboratory designs which undermine their external validity. Different samples, types of stimuli, measurement modes, and research contexts may violate transportability in that they produce varying rates of observed behaviors and—more worryingly for experimental research—heterogeneous treatment effects. In four studies, we assessed each dimension’s importance for establishing transportability.

We demonstrated that a common class of laboratory designs—interactive games measuring fairness, trust, and reciprocity—easily violates the transportability of percentage rates of observed behavior: First, synchronized lab implementations revealed substantial differences in elicited behavior between two locally recruited student-subject pools (Study 1). This cross-location gap persists in alternative online implementations (Study 2), which shut off potential experimenter effects and differences in labs’ physical appearances. One may thus speculate about regional idiosyncrasies bringing about specific patterns of behavior that jeopardize *pool generalizability*. Second, we find much higher rates of prosocial behavior among a broader population sample (Study 3), indicating a lack of *sample generalizability*, as results yielded from student participants do not transport to a more representative population. Third, in a replication at MTurk (Study 4) we find rates of prosocial behavior even lower than in our student samples. This clearly rejects *context generalizability* of quantitative results.

Even when keeping socio-demographics constant, data collected from the most frequently used participant groups, students and crowdworkers, differ significantly from those obtained from the broader population—and altruistic behavior as measured in the Dictator Game (DG) proved to be particularly sensitive to changes in units and settings. We chose stake levels typical for the respective participant pool. As a side effect, we cannot fully rule out the

possibility that differences across samples and contexts may be partly due to differences in monetary incentives. Given the well-documented finding that specific sizes of positive stakes have negligible effects in interactive games of fairness, trust, and reciprocity (Carpenter et al. 2005; Keuschnigg et al. 2016), it is highly unlikely that stake differences drive our results. In fact, we find the lowest level of prosocial behavior in the setup providing the smallest stakes (Study 4)—a finding that runs counter the idea that prosociality decreases in stake sizes.

Our unstable quantitative results indicate that preference parameters (such as “prosociality”) measured in laboratory designs can not be transported to other populations, and their use in establishing descriptive results about “human nature” is questionable. The heterogeneity in elicited behavior that we found for decision situations targeting prosociality presumably also affects laboratory studies using other types of decision situations. Hence, interpretations of marginal totals obtained in laboratory research remain descriptive and studies reporting an intervention’s consequence in absolute terms risk describing only highly local results. However, nobody in the social sciences would expect a volunteer sample of, say, student respondents to generate survey data identical to a random population sample. The local bound of descriptive results is thus not an exclusive feature of laboratory designs but mainly a sampling issue.

Against this cautionary backdrop, our results sustain an optimistic view for the external validity of theory-driven experiments focusing on the identification of causal effects. Qualitative results, in our case the finding that transfers in the Trust Game (TG) on average exceed DG allocations, are remarkably robust across samples, measurement modes, and research contexts. The problem of unstable results re-emerges, however, if experimenters estimate treatment effects from contrasting with an unstable control condition. In our studies, differences in mean transfers between DG and TG vary as altruistic decisions in DG interact with both the characteristics of the sample under consideration and the specific setting. If control conditions provide unstable measures like DG, point estimates of treatment effects may be seriously biased. Heterogeneous treatment effects then stem from an unstable control condition rather than from heterogeneous responses to the treatment itself.

Similarly important for practitioners, we find that specific implementations of a laboratory study do not distort its results. For our physical labs, we find no evidence for *anonymity effects* in data collection—although decisions concerning prosocial behavior should be particularly liable to social desirability bias. If we increase participant anonymity, rates of elicited behavior do not differ significantly from conditions lacking specific anonymization measures. Our results are thus in line with Barmettler et al. (2012), who question the necessity of complicated anonymization procedures in social science laboratories. Reactivity may still drive behavior, but we find no effect within the spectrum of anonymity precautions typically used in experiments. This suggests comparability of results from laboratory studies differing in this respect. Finally, we find full support for *mode generalizability*. Keeping participation effort constant across parallel lab and online sessions, participants from two student-subject pools generated similar data, irrespective of participating in the lab or online. Taking into account laboratory studies’ weak generalizability to a broader, more representative population, we believe that online experiments can serve as a sorely needed complement to laboratory designs in the social sciences.

To conclude, successful and meaningful laboratory research in sociology—just as in any other empirical discipline—requires joint efforts of ongoing replication. We can only regard laboratory results as well-established facts after their successful cross-validation, ideally in studies using complementary samples and designs. This also holds for our results, which we hope will be replicated in future studies using alternative decision situations frequently used in social science lab research.

Notes

1. Camerer (2003), Cooper and Kagel (2016), and Fehr and Gintis (2007) provide overviews of laboratory designs and findings related to prosociality. Fréchette (2016), Fréchette and Schotter (2015, part 4), and Galizzi and Navarro-Martínez (2018) review the literature on the external validity of experimental game theory.
2. Belot et al. (2015) rely on ad-hoc sampling of residents from the community surrounding the University of Oxford, United Kingdom. Cappelen et al. (2015) use quota sampling in municipalities surrounding Bergen, Norway. Falk et al. (2013) use a representative sample of the city of Zurich, Switzerland. Anderson et al. (2013) feature a sophisticated random sample representative of the Danish population. To homogenize participation effort, they conducted sessions at multiple locations across the country. Other frequently cited variants of this design are less instructive, as they fail to standardize experimental conditions fully across samples (Bellemare and Kröger 2007) or lack student comparison groups (Fehr et al. 2002).
3. Falk et al. (2013) find that differences across samples disappear after conditioning on socio-demographic composition; this implies that students are not less prosocial *per se*, as non-student participants with comparable backgrounds show similar rates of selfish behavior. The majority of studies, however, find that sample differences resist conditioning (Anderson et al. 2013; Belot et al. 2015; Cappelen et al. 2015).
4. Actual matching occurred only after elicitation of behavioral data to avoid waiting time: To determine individual payoff at the end of the session, we randomly selected one of the decisions by each finalist to pair with a complementing decision by another participant.
5. Our overall DG result (42.2%) clearly differs from the mean allocation of 28.4% reported in Engel’s (2011) meta-analysis of more than 600 DG results ($t=32.44$, $p<.001$). Trustors, on the other hand, invest 52.6% of their endowment on average. Although statistically different ($t=4.48$, $p<.001$), this rate approximates the mean investment of 50.2% reported in Johnson’s and Mislin’s (2011) meta-analysis of 161 TG results. We further find sequence effects—transfers in DG and TG are significantly larger if the respective game is played first (see Appendix A2)—but we randomized sequences such that they do not explain differences in mean transfers between games.
6. We find support for this and the following results in replications using the proportion of nonzero transfers by dictators and trustors (see Appendix A4) and using data on second-

mover behavior in TG as well as, for Studies 1 and 2, data obtained from an additional Ultimatum Game (see Appendix A5).

7. In Figures 1 and 3, we pooled these three laboratory conditions to display average results.
8. Contrasting online results to the low-anonymity condition, we yield $t=.76$, $p=.446$ in DG and $t=1.77$, $p=.078$ in TG. Comparison to the standard-anonymity condition yields $t=.30$, $p=.765$ in DG and $t=.56$, $p=.578$ in TG and to the high-anonymity condition $t=.84$, $p=.400$ in DG and $t=1.12$, $p=.263$ in TG.

References

- Alevy, Jonathan E., Michael S. Haigh, and John A. List 2007. "Information Cascades: Evidence from a Field Experiment with Financial Market Professionals." *Journal of Finance* 62(1):151-80.
- Amir, Ofra, David G. Rand, and Ya'akov Kobi Gal 2012. "Economic Games on the Internet: The Effect of \$1 Stakes." *PLoS ONE* 7(2):e31461.
- Anderson, Jon, Stephen V. Burks, Jeffrey Carpenter, Lorenz Götte, Karsten Maurer, Daniele Nosenzo, Ruth Potter, Kim Rocha, and Aldo Rustichini 2013. "Self-selection and Variations in the Laboratory Measurement of Other-regarding Preferences Across Subject Pools: Evidence From One College Student and Two Adult Samples." *Experimental Economics* 16(2):170-89.
- Barmettler, Franziska, Ernst Fehr, and Christian Zehnder 2012. "Big Experimenter is Watching You! Anonymity and Prosocial Behavior in the Laboratory." *Games and Economic Behavior* 75(1):17-34.
- Bellemare, Charles and Sabine Kröger 2007. "On Representative Social Capital." *European Economic Review* 51(1):183-202.
- Belot, Michele, Raymond Duch, and Luis Miller 2015. "A Comprehensive Comparison of Students and Non-students in Classic Experimental Games." *Journal of Economic Behavior & Organization* 113:26-33.
- Benz, Matthias and Stephan Meier 2008. "Do People Behave in Experiments as in the Field? Evidence from Donations." *Experimental Economics* 11(3):268-81.
- Beramendi, Pablo, Raymond M. Duch, and Akitaka Matsuo 2016. "Comparing Modes and Samples in Experiments: When Lab Subjects Meet Real People." SSRN Research Paper 2840403.
- Berg, Joyce E., John Dickhaut, and Kevin McCabe 1995. "Trust, Reciprocity, and Social History." *Games and Economic Behavior* 10(1):122-42.
- Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3):351-68.
- Bicchieri, Cristina 2006. *The Grammar of Society: The Nature and Dynamics of Social Norms*. New York: Cambridge University Press.
- Bock, Olaf, Ingmar Baetge, and Andreas Nicklisch 2014. "hroot: Hamburg Registration and Organization Online Tool." *European Economic Review* 71:117-20.
- Bolle, Friedel 1990. "High Reward Experiments without High Expenditure for the Experimenter?" *Journal of Economic Psychology* 11(2):157-67.
- Brandts, Jordi, Tatsuyoshi Saijo, and Arthur Schram 2004. "How Universal is Behavior? A Four Country Comparison of Spite and Cooperation in Voluntary Contribution Mechanisms." *Public Choice* 119(3):381-424.
- Camerer, Colin F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. New York: Sage.
- Camerer, Colin F. and Ernst Fehr 2004. "Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists." Pp. 55-95 in *Foundations of Human Sociality*, edited by J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, and H. Gintis. Oxford, UK: Oxford University Press.

- Camerer, Colin F. and Robin M. Hogarth 1999. "The Effects of Financial Incentives in Experiments: A Review and Capital-labor-production Framework." *Journal of Risk and Uncertainty* 19(1-3):7-42.
- Campbell, Donald T. and Julian C. Stanley 1963. *Experimental and Quasi-experimental Designs for Research*. Chicago, IL: Rand-McNally.
- Cappelen, Alexander W., Knut Nygaard, Erik O. Sorensen, and Bertil Tungodden 2015. "Social Preferences in the Lab: A Comparison of Students and a Representative Population." *Scandinavian Journal of Economics* 117(4):1306-26.
- Carpenter, Jeffrey, Eric Verhoogen, and Stephen Burks 2005. "The Effect of Stakes in Distribution Experiments." *Economics Letters* 86(3):393-8.
- Chandler, Jesse, Pam Mueller, and Gabriele Paolacci 2014. "Nonnaïveté Among Amazon Mechanical Turk Workers: Consequences and Solutions for Behavioral Researchers." *Behavioral Research Methods* 46(1):112-30.
- Chang, Andrew C. and Phillip Li 2015. "Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say 'Usually Not'." *Finance and Economics Discussion Series* 2015-083.
- Clifford, Scott and Jennifer Jerit 2014. "Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory and Online Studies." *Journal of Experimental Political Science* 1(2):120-31.
- Cooper, David J. and John H. Kagel 2016. "Other-regarding Preferences: A Selective Survey of Experimental Results." Pp. 217-89 in *The Handbook of Experimental Economics, Vol. 2.*, edited by J. Kagel and A. Roth. Princeton, NJ: Princeton University Press.
- Coppock, Alexander and Donald P. Green 2015. "Assessing the Correspondence between Experimental Results Obtained in the Lab and Field: A Review of Recent Social Science Research." *Political Science Research and Methods* 3(1):113-31.
- Cronbach, Lee J. 1982. *Designing Evaluations of Educational and Social Programs*. San Francisco, CA: Jossey-Bass.
- Crump, Matthew J.C., John V. McDonnell, and Todd M. Gureckis 2013. "Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research." *PLoS ONE* 8(3):e57410.
- Druckman, James N. and Cindy D. Kam 2011. "Students as Experimental Participants." Pp. 41-57 in *The Cambridge Handbook of Experimental Political Science*, edited by J. Druckman, D. Green, J. Kuklinski, and A. Lupia. Cambridge: Cambridge University Press.
- Dunning, Thad 2012. *Natural Experiments in the Social Sciences: A Design-based Approach*. Cambridge: Cambridge University Press.
- Elster, Jon 2007. *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*. Cambridge, MA: Cambridge University Press.
- Engel, Christoph 2011. "Dictator Games: A Meta Study." *Experimental Economics* 14(4):583-610.
- Englmaier, Florian and Georg Gebhardt 2016. "Social Dilemmas in the Laboratory and in the Field." *Journal of Economic Behavior & Organization* 128:85-96.
- Falk, Armin and James Heckman 2009. "Lab Experiments Are a Major Source of Knowledge in the Social Sciences." *Science* 326(5952):535-8.
- Falk, Armin, Stephan Meier, and Christian Zehnder 2013. "Do Lab Experiments Misrepresent Social Preferences? The Case of Self-selected Student Samples." *Journal of the European Economic Association* 11(4):839-52.
- Fehr, Ernst, Urs Fischbacher, Bernhard von Rosenbladt, Jürgen Schupp, and Gert G. Wagner 2002. "A Nation-wide Laboratory: Examining Trust and Trustworthiness by Integrating Behavioral Experiments into Representative Surveys." *Schmollers Jahrbuch* 122:519-42.
- Fehr, Ernst and Herbert Gintis 2007. "Human Motivation and Social Cooperation: Experimental and Analytical Foundations." *Annual Review of Sociology* 33:43-64.
- Fehr, Ernst and John A. List 2004. "The Hidden Costs and Returns of Incentives: Trust and Trustworthiness Among CEOs." *Journal of the European Economic Association* 2(5):743-71.

- Franzen, Axel and Sonja Pointner 2012. "Anonymity in the Dictator Game Revisited." *Journal of Economic Behavior & Organization* 81:74-81.
- Franzen, Axel and Sonja Pointner 2013. "The External Validity of Giving in the Dictator Game: A Field Experiment Using the Misdirected Letter Technique." *Experimental Economics* 16(2):155-69.
- Fréchette, Guillaume R. 2015. "Laboratory Experiments: Professionals versus Students." Pp. 360-90 in *The Handbook of Experimental Economic Methodology*, edited by G. Fréchette and A. Schotter. New York: Oxford University Press.
- Fréchette, Guillaume R. 2016. "Experimental Economics Across Subject Populations." Pp. 435-80 in *The Handbook of Experimental Economics, Vol. 2.*, edited by J. Kagel and A. Roth. Princeton, NJ: Princeton University Press.
- Fréchette, Guillaume R. and Andrew Schotter 2015. *The Handbook of Experimental Economic Methodology*. New York: Oxford University Press.
- Freese, Jeremy 2007. "Replication Standards for Quantitative Social Science: Why Not Sociology?" *Sociological Methods and Research* 36(2):153-72.
- Galizzi, Matteo M. and Daniel Navarro-Martínez 2018. "On the External Validity of Social-preference Games: A Systematic Lab-field Study." *Management Science*, Article in Advance.
- Gerber, Alan S. and Donald P. Green 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York: Norton.
- Glaeser, Edward L., David I. Laibson, Jose A. Scheinkman, and Christine L. Soutter 2000. "Measuring Trust." *Quarterly Journal of Economics* 115(3):811-46.
- Gosling, Samuel D., Carson J. Sandy, Oliver P. John, and Jeff Potter 2010. "Wired But Not WEIRD: The Promise of the Internet in Reaching More Diverse Samples." *Behavior and Brain Science* 33(2-3):34-35.
- Harrison, Glenn W. and John A. List 2004. "Field Experiments." *Journal of Economic Literature* 42(4):1009-55.
- Henrich, Joseph, Richard Boyd, Samuel Bowles, Colin F. Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath 2001. "In Search of Homo Economicus: Behavioral Experiments in 15 Small-scale Societies." *American Economic Review* 91(2):73-8.
- Henrich, Joseph, Steven J. Heine, and Ara Norenzayan 2010. "The Weirdest People in the World?" *Behavioral and Brain Sciences* 33(2-3):1-23.
- Hergueux, Jérôme and Nicolas Jacquemet 2015. "Social Preferences in the Online Laboratory: A Randomized Experiment." *Experimental Economics* 18(2):251-83.
- Hoffman, Elizabeth, Kevin A. McCabe, and Vernon L. Smith 1996. "Social Distance and Other-regarding Behavior in Dictator Games." *American Economic Review* 86(3):653-60.
- Horton, John J., David G. Rand, and Richard J. Zeckhauser 2011. "The Online Laboratory: Conducting Experiments in a Real Labor Market." *Experimental Economics* 14(3):399-425.
- Hox, Joop, Edith de Leeuw, and Thomas Klausch 2017. "Mixed-mode Research: Issues in Design and Analysis." Pp. 511-30 in *Total Survey Error in Practice*, edited by P.P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L.E. Lyberg, N.C. Tucker, and B.T. West. Hoboken, NJ: Wiley.
- Ipeirotis, Panagiotis G. 2017. MTurk Tracker. <http://demographics.mturk-tracker.com/#/countries/all>.
- Jackson, Michelle and David R. Cox 2013. "The Principles of Experimental Design and Their Application in Sociology." *Annual Review of Sociology* 39:27-49.
- Johnson, Noel D. and Alexandra A. Mislin 2011. "Trust Games: A Meta-analysis." *Journal of Economic Psychology* 32(5):865-89.
- Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler 1986. "Fairness and the Assumptions of Economics." *Journal of Business* 59(4):285-300.
- Kessler, Judd and Lise Vesterlund 2013. "The External Validity of Laboratory Experiments: The Misleading Emphasis on Quantitative Effects." Pp. 391-406 in *The Handbook of Experimental Economic Methodology*, edited by G. R. Fréchette and A. Schotter. New York: Oxford University Press.

- Keuschnigg, Marc, Felix Bader, and Johannes Bracher 2016. "Using Crowdsourced Online Experiments to Study Context-dependency of Behavior." *Social Science Research* 59:68-82.
- Kocher, Martin, Todd Cherry, Stephan Kroll, Robert J. Netzer, and Matthias Sutter 2008. "Conditional Cooperation on Three Continents." *Economic Letters* 101(3):175-8.
- Levitt, Steven D. and John A. List 2007. "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?" *Journal of Economic Perspectives* 21(2):153-74.
- Martin, Michael W. and Jane Sell 1979. "The Role of the Experiment in the Social Sciences." *Sociological Quarterly* 20(4):581-90.
- Morgan, Stephen L. and Christopher Winship 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, 2nd ed. New York: Cambridge University Press.
- Open Science Collaboration 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349(6251):943-51.
- Pearl, Judea and Elias Bareinboim 2014. "External Validity: From Do-calculus to Transportability Across Populations." *Statistical Science* 29(4):579-95.
- Peterson, Robert A. 2001. "On the Use of College Students in Social Science Research: Insights From a Second-order Meta-analysis." *Journal of Consumer Research* 28(3):450-61.
- Potters, Jan and Frans van Winden 2000. "Professionals and Students in a Lobbying Experiment Professional Rules of Conduct and Subject Surrogacy." *Journal of Economic Behavior & Organization* 43:499-522.
- Rand, David G. 2012. "The Promise of Mechanical Turk: How Online Labor Markets Can Help Theorists Run Behavioral Experiments." *Journal of Theoretical Biology* 299:172-79.
- Rand, David G., Alexander Peysakhovich, Gordon T. Kraft-Todd, George E. Newman, Owen Wurzbacher, Martin A. Nowak, and Joshua D. Greene 2014. "Social Heuristics Shape Intuitive Cooperation." *Nature Communications* 5:3677.
- Rauhut, Heiko and Fabian Winter 2010. "A Sociological Perspective on Measuring Social Norms by Means of Strategy Method Experiments." *Social Science Research* 39(6):1181-94.
- Reips, Ulf-Dietrich 2002. "Standards for Internet-based Experimenting." *Experimental Psychology* 49(4):243-56.
- Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir 1991. "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study." *American Economic Review* 81(5):1068-95.
- Rubin, Donald B. 2008. "For Objective Causal Inference, Design Trumps Analysis." *Annals of Applied Statistics* 2(3):808-40.
- Schram, Arthur 2005. "Artificiality: The Tension Between Internal and External Validity in Economic Experiments." *Journal of Economic Methodology* 12(2):225-37.
- Shadish, William R., Thomas D. Cook, and Donald T. Campbell 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.
- Thye, Shane R. 2014. "Logical and Philosophical Foundations of Experimental Research in the Social Sciences." Pp. 53-82 in *Laboratory Experiments in the Social Sciences*, edited by M. Webster, Jr. and J. Sell. Burlington, MA: Academic Press.
- Webster, Murray, Jr. and Jane Sell 2014. "Why Do Experiments?" Pp. 5-21 in *Laboratory Experiments in the Social Sciences*, edited by M. Webster, Jr. and J. Sell. Burlington, MA: Academic Press.
- Willer, David and Henry A. Walker 2007. *Building Experiments: Testing Social Theory*. Stanford, CA: Stanford University Press.
- Zelditch, Morris, Jr. 2014. "Laboratory Experiments in Sociology." Pp. 183-97 in *Laboratory Experiments in the Social Sciences*, edited by M. Webster, Jr. and J. Sell. Burlington, MA: Academic Press.

Figures and Table

Figure 1: Quantitative results

Note: Shaded bars show unconditional means of first-mover transfers in the Dictator Game (DG) and the Trust Game (TG), respectively. Blank bars represent conditional means obtained from OLS regressions keeping underlying socio-demographics constant. We include 95% confidence intervals and seven pairwise comparisons (t -tests). *** $p < .001$, ** $p < .01$, * $p < .05$, n.s.=non-significant.

Figure 2: Anonymity conditions in the laboratory

Note: Shaded bars show unconditional means of first-mover transfers in the Dictator Game (DG) and the Trust Game (TG), respectively. Blank bars represent conditional means obtained from OLS regressions keeping underlying sociodemographics constant. We include 95% confidence intervals. All pairwise comparisons between anonymity conditions are non-significant.

Figure 3: Qualitative results

Note: Shaded bars show conditional means of first-mover transfers in the Dictator Game (DG). Blank bars on top represent conditional means in the Trust Game (TG). 95% confidence intervals, here, indicate significance of the within-subject difference DG–TG (paired t -tests).

Table 1. Study details.

Study	Location	Participants	Data collection	N	# Prior experiments	% Without experience	Endowment DG	TG	
1	<i>Pool generalizability</i> Parallel lab sessions with newly recruited participant pools	Leipzig Munich	Local students	Apr 21–Jun 7, 2016	362 351	1.6	70.4	10.0€	5.0€
2	<i>Mode generalizability</i> Parallel online sessions with newly recruited participant pools	Leipzig Munich	Local students	Apr 21–Jun 7, 2016	122 115	1.0	78.5	10.0€	5.0€
3	<i>Sample generalizability</i> Nationwide online experiment	Germany	Representative of German-born population regarding gender, age (18–69), region	Jun 10–27, 2016	1,223	1.2	84.2	5.0€	2.5€
4	<i>Context generalizability</i> Replication in an online labor market	MTurk	MTurk workers from the United States and India	Mar 4–Jun 3, 2017	491	65.5	49.5	2.0\$	1.0\$

Note: # Prior experiments refers to the average number of incentivized experiments subjects had taken prior to participation. Studies 1 and 2, at each location, draw on the same local student pool. We used the web-based software hroot (Bock, Baetge, and Nicklisch 2014) to randomize invitations to lab and online sessions. Endowments refer to euros (Studies 1–3) or US dollars (Study 4).

Appendix

A1 Socio-demographics across Samples and Control Variables

Table A1 summarizes the different socio-demographic backgrounds of our participants as well as relevant control variables capturing properties of our design potentially affecting elicited behavior. We include these characteristics in our calculations of conditional means (see Appendix A2 for the full models). We indicate binary measures with (0,1); all other measures are continuous. We test, further, for significant differences in sample composition, focusing on the seven pairwise comparisons that are crucial to our main analysis.

Gender. Women are broadly overrepresented in both student pools in Leipzig and Munich. The nationwide sample, in which gender is one of the criteria for stratified sampling, has a balanced sex ratio. Women, however, make up only a third of the MTurk sample.

Age. On average, student participants are 22 years old, the broader population 45, and crowdworkers 33. Because age often correlates non-linearly with elicited behavior in the lab (see, for an overview, Fréchette 2016), we categorize age into intervals 18–22, 23–29, 30–45, and ≥ 46 for inclusion in our regressions (see Appendix A2).

Education. Participants in our student pools differ with respect to study fields. Among Leipzig students, most pursue a major in humanities (e.g., cultural studies, literature, pedagogy). In Munich, most major in other programs such as social sciences, law, or STEM fields. We particularly differentiate between students of economics and of the humanities. Prior studies have shown that the former behave more selfishly in incentivized games (Etzioni 2015). We use all other study fields as our reference category. In the nationwide sample, 60% have at least a high-school diploma and 9% are currently enrolled in a study program. Among crowdworkers, 90% have at least secondary education and 12% are currently students.

Employment. For student subjects, we code 1 if they have a side job (50% in Leipzig and 63% in Munich). In the non-student samples, we code 1 for the fully-employed. This is true for 58% in the nationwide sample and for 69% of crowdworkers.

Income. We measure monthly disposable income using the PPP\$-adjusted household income divided by the square root of household size. On average, students are significantly more affluent in Munich (PPP\$1,282) than in Leipzig (PPP\$870). Both groups have substantially less income at their disposal than the average German (PPP\$2,727). MTurk participants take a middle position with an average income of PPP\$1,562. In our regressions, we include a log transformation of this variable.

Parenthood. A number of observational studies find that parents differ in prosocial behavior compared to individuals without children (see Wiepking and Bekkers 2012 for an overview). Parenthood is rare among students (1–2%). In the nationwide sample, 57% have at least one child. This rate is 43% for crowdworkers.

Experience. Prior experience with laboratory games varies considerably across studies. Both student subjects and members of the broader population on average had only taken 1–3 social experiments prior to participation. For MTurk workers, this number amounts to 66 on average.

Understanding. For each game, participants had to answer two control questions following their decisions (see Appendix A6 for examples). Monitoring the understanding of instructions we allowed for repeated submissions, each time informing participants whether their entries were correct or incorrect. Our binary indicator of misunderstanding takes the value 1 for each subject providing more than three incorrect entries in either the Dictator Game (DG) or the Trust Game (TG). 85% of those coded 1 struggled with questions on TG. The fraction of participants scoring 1 is lowest among students (15%), but relatively high in both the broader population (43%) and the MTurk sample (34%).

Surroundings during online participation. We collected self-reported data on the physical and social surrounding during online participation. Most online subjects participated from their home, although this rate is significantly smaller for the two student samples (62%) than for either the broader population (85%) or the MTurk sample (87%). Only a few reported that others were observing them while making their decisions: Rates are 9% for both students and crowdworkers, and 6% for the broader population.

Decision sequence. To account for sequence effects, we include in our regressions a binary variable indicating whether the current decision was not the focal subject’s first. Note that we randomized participants to specific sequences of games and first- and second-mover roles.

The tabulation finally demonstrates effective randomization of student participants to modes of data collection (see columns “a vs c” and “b vs d”). With the exception of age ($p = .026$ for the Munich pool), differences in socio-demographics between lab and online participants are non-significant.

Table A1. Socio-demographics across samples.

		Study 1 Lab			Study 2 Online				
Variable	Measure	(a)	(b)	a vs b	(c)	(d)	c vs d	a vs c	b vs d
		Leipzig	Munich	<i>p</i>	Leipzig	Munich	<i>p</i>	<i>p</i>	<i>p</i>
Gender	Female (0,1)	0.70	0.66	.279	0.68	0.70	.690	.746	.353
Age	Years since birth	22.34	22.21	.553	22.31	21.53	.037	.920	.026
Students' major	Economics (0,1)	0.12	0.15	.209	0.10	0.12	.568	.524	.419
	Humanities (0,1)	0.47	0.37	.006	0.53	0.43	.132	.250	.227
	Other (0,1)	0.41	0.48	.061	0.37	0.44	.244	.433	.513
Employment	Side job (0,1)	0.51	0.63	.001	0.48	0.64	.009	.498	.789
Income	Monthly disposable income	879.85	1290.88	<.001	841.25	1253.13	<.001	.177	.268
Parenthood	≥1 Child (0,1)	0.02	0.01	.432	0.01	0.01	.967	.218	.606
Experience	Number of prior lab experiments	0.52	2.61	<.001	0.34	1.64	.002	.286	.062
Control questions	>3 Mistakes in DG or TG (0,1)	0.13	0.17	.082	0.12	0.16	.459	.905	.663
Online participation	At home (0,1)				0.67	0.57	.120		
	Observed by others (0,1)				0.06	0.13	.056		
Decision sequence	DG not first decision (0,1)	0.68	0.64	.278	0.66	0.68	.714	.632	.464
	TG trustor not first decision (0,1)	0.83	0.83	.775	0.86	0.86	.996	.477	.366
<i>N</i>		362	351		122	115			

		Study 3 Nationwide		Study 4 MTurk		
Variable	Measure	(e)	c+d vs e	(f)	c+d vs f	e vs f
			<i>p</i>		<i>p</i>	<i>p</i>
Gender	Female (0,1)	0.49	<.001	0.35	<.001	<.001
Age	Years since birth	44.61	<.001	32.98	<.001	<.001
Education	≥High-school diploma (0,1)	0.60		0.90		<.001
	Currently student (0,1)	0.09		0.12		.035
Employment	Fully employed (0,1)	0.58	.534	0.69	<.001	<.001
Income	Monthly disposable income	2727.03	<.001	1561.55	<.001	<.001
Parenthood	≥1 Child (0,1)	0.57	<.001	0.43	<.001	<.001
Experience	Number of prior lab experiments	1.18	.410	65.52	.005	.005
Control questions	>3 Mistakes in DG or TG (0,1)	0.43	<.001	0.34	<.001	<.001
Online participation	At home (0,1)	0.85	<.001	0.87	<.001	.321
	Observed by others (0,1)	0.06	.128	0.09	.970	.034
Decision sequence	DG not first decision (0,1)	0.66	.955	0.50	<.001	<.001
	TG trustor not first decision (0,1)	0.67	<.001	0.50	<.001	<.001
<i>N</i>		1,223		491		

Note: We report means and *p*-values from *t*-tests for between-sample differences.

A2 Correlates of Elicited Behavior

The conditional means reported in the main text are predicted values ($\hat{Y}$) for each Study obtained from pooled OLS regressions. Table A2 displays these models. The dependent variable is first-mover behavior in either the DG or TG. We allowed individual transfers in each game to be multiples of 10% of the endowment (including 0%). Hence, $Y \in [0, 100]$. Because our dependent variables are both left and right censored we also ran tobit regressions, which yield comparable results. We repeat the predicted values mentioned in the text at the top of the table. The lower part shows how elicited behavior correlates with socio-demographic characteristics and how properties of our design affect subjects' choices. The reported coefficients rest on non-experimental variation and remain descriptive, even when controlling for all known confounders.

First, there are apparent differences between games regarding the potential influence of individual characteristics: Dictator behavior appears amenable to socio-demographic influence. TG, on the other hand, remains less affected by individual characteristics. Differences in model determination (R^2) reflect this fact.

Second, we find noteworthy results for single variables. Women, although behaving similarly to men in DG, are significantly less trusting in TG ($p < .001$). Compared to subjects aged 18–22, older participants share considerably less in both DG and TG. For the latter game, this difference is only significant for those aged 46 and older ($p = .011$). With respect to students' study field, our results are consistent with prior evidence (Etzioni 2015), indicating that economics majors behave more selfishly in incentivized games: Compared to majors in other areas, they give almost 11.0 percentage points less of their endowment in DG ($p < .001$) and transfer 8.3 percentage points less in TG ($p = .010$). Students of the humanities, in contrast, give 2.7 percentage points more in DG ($p = .018$) than their counterparts from other disciplines. Affluent participants, in turn, are less giving: A 10% higher disposable income associates with 0.2 percentage points lower dictator allocations ($p = .013$). Again, there is no such association in TG. Experimental experience, finally, correlates negatively with dictator allocations ($p < .001$). This finding is in line with prior results suggesting that frequent participation in social experiments provides the opportunity to calibrate one's responses and to arrive at behavior more consistent with the standard economic theory (Levitt and List 2007; Rand et al. 2014). In our data, however, this association is extremely weak (with each prior participation average allocation decreases by 0.005 percentage points) and absent for prosocial behavior as measured in TG.

Third, we show how specific properties of our design correlate with participants' choices. Subjects having difficulties in understanding the instructions exhibit considerably more prosocial behavior in DG ($p < .001$) but not in TG ($p = .676$). We address the crucial issue of comprehension in Appendix A3. Here, we stress the importance of controlling for differences in understanding to attain conditional means unaffected by some participants' faulty responses. Online subjects being observed by others during participation feature higher rates of prosocial behavior than those making their decisions in isolation ($p = .024$ in DG, $p = .046$ in TG). Finally, our analyses control for the sequence of games and roles (note that this variable is a randomized treatment). Not making the focal decision at first has a significant negative effect on prosocial behavior in both DG ($p = .034$) and TG ($p < .001$). This effect is most pronounced for placing trust, where having had previous decisions reduces sharing by

Table A2. Correlates of elicited behavior

		DG		TG	
Predicted means		$\hat{Y}$		$\hat{Y}$	
Study 1	Leipzig (lab)	42.296		52.379	
	Munich (lab)	36.377		48.607	
Study 2	Leipzig (online)	42.938		48.261	
	Munich (online)	35.814		48.554	
Study 3	Nationwide	47.691		58.808	
Study 4	MTurk	33.721		42.371	
		β	t	β	t
Gender	Female	0.583	.77	-5.674***	5.01
Age	18–22 (reference)				
	23–29	-4.445***	4.13	-1.639	.99
	30–45	-4.136*	2.41	-4.387	1.80
	≥46	-3.982*	2.25	-6.585*	2.54
Students' major	Other (reference)				
	Economics	-10.890***	5.23	-8.325*	2.57
	Humanities	2.695*	2.37	-1.790	.95
Education	≥High-school diploma	0.263	0.27	-0.398	.28
	Currently student	-2.721	1.31	1.482	.56
Income	log(Monthly disposable income)	-2.018*	2.50	-0.714	.76
Employment	Fully employed	-0.540	.69	-1.966	1.70
Parenthood	≥1 Child	1.811	1.68	0.329	1.22
Experience	Number of prior lab experiments	-0.005***	6.34	-0.003	1.25
Control questions	>3 Mistakes in DG or TG	4.350***	5.30	-0.473	.42
Online participation	At home	-1.429	1.34	-0.500	.31
	Observed by others	4.248*	2.26	4.413*	2.00
Decision sequence	Current decision not first	-1.575*	2.12	-4.504***	3.88
N		2,664		2,664	
R^2		0.139		0.059	

Note: OLS regressions on first-mover behavior in the Dictator Game (DG) and the Trust Game (TG), respectively. We report unstandardized coefficients and t -values calculated from robust standard errors. *** $p < .001$, ** $p < .01$, * $p < .05$.

4.5 percentage points on average.

A3 Comprehension

In our studies, we place participants into decision situations in which making mistakes and showing prosocial behavior are, on average, confounded (Andreoni 1995; Recalde, Riedl, and Vesterlund 2018): Ambiguous individual responses in the middle of the choice set indicate sharing rates of 50% of the endowment. The same holds for the average over many random responses on a 0–100% continuum. Hence, “mistakes and non-standard preferences move in the same direction (away from predictions for perfectly optimizing and selfish agents)” (Fréchette 2016:472).

To test whether difficulties in understanding drive prosocial choices in our nationwide

Figure A1: Treatment effects of time pressure

Note: Blank bars show unconditional means of first-mover transfers under time delay in the Dictator Game (DG) and the Trust Game (TG), respectively. Shaded bars represent unconditional means under time pressure. We include 95% confidence intervals and two pairwise comparisons (t -tests). n.s.=nonsignificant.

sample, we randomized participants in Study 3 into a time-limit treatment. Under *time pressure*, we asked 50% of our participants, after they had read their instructions, to take their decisions within 10 seconds in DG and within 15 seconds in TG, adjusting time pressure for the complexity of the decision situation. 17.6% of exposed subjects did not comply with our time limit; to avoid selection bias (cf., Tinghög et al. 2013), we allowed entries outside the time frame and include non-compliers in our analysis. The other half faced *time delay*, under which they could only take each decision after a waiting time of 20 seconds.

If differences in comprehension explain prosocial behavior in our broader population sample, behavior should map students' behavior more closely under the time-delay condition, permitting active deliberation and reconsideration of instructions. Under time pressure, one can plausibly assume that participants will deliberate less and make more mistakes. The latter should evoke random responses and thus move sample averages toward the middle of the choice set.

We find no statistically significant effect of the time-pressure treatment in either DG or TG (Figure A1), suggesting that differences in comprehension do not explain higher rates of prosocial behavior among the broader population.

A4 Replication: Proportions of Nonzero Transfers

Differences between our samples' mean transfers in DG and TG partly stem from the proportions of nonzero transfers. In contrast to the standard economic theory, which predicts zero transfers in both decision situations (e.g., Engel 2011, Johnson and Mislin 2011), 90.5% of our subjects shared positive amounts in DG and 92.8% invested in TG. Confirming our results, we find behavioral differences similar to our main findings when, alternatively, focusing on proportions of nonzero transfers (Figure A2). We report p -values of two-sided z -tests.

Figure A2: Proportions of Nonzero Transfers

Note: Shaded bars show unconditional proportions of nonzero first-mover transfers in the Dictator Game (DG) and the Trust Game (TG), respectively. Blank bars represent conditional proportions obtained from logistic regressions keeping underlying socio-demographics constant. We include 95% confidence intervals and seven pairwise comparisons (z -tests). *** $p < .001$, ** $p < .01$, * $p < .05$, n.s.=nonsignificant.

Among the Leipzig students who participated in lab sessions, 94.5% shared positive amounts in DG; this rate is 84.3% in Munich. This difference between student pools remains when controlling for socio-demographics ($z=3.23$, $p=.001$). In TG, the difference is smaller (95.6% in Leipzig and 89.7% in Munich transferred positive amounts), but—in contrast to the results on mean transfers—the difference remains significant under conditioning on socio-demographics ($z=2.10$, $p=.036$). Corroborating our finding for mean transfers, pool differences in the rates of nonzero transfers are larger in DG.

Again, we find no mode effect. Within locations, rates of nonzero transfers in online sessions are similar to and not significantly different from those in the laboratory (DG in Leipzig $z=0.14$, $p=.891$ and in Munich $z=0.24$, $p=.810$; TG in Leipzig $z=0.09$, $p=.925$ and in Munich $z=0.24$, $p=.814$).

The broader population sample, again, shows the highest rates of prosocial behavior in both decision situations—even when holding socio-demographics constant: 97.4% transfer a positive amount in DG (compared to students online 88.6%; $z=3.48$, $p<.001$) and 98.2%

in TG ($z=3.68$, $p<.001$). Like before, our implementation at MTurk returns the lowest rates of prosocial behavior in both decision situations: Adjusted for differences in socio-demographics, only 78.2% transfer a positive amount in DG and 81.8% in TG. In DG, these rates are significantly lower than those obtained in both the broader population ($z=9.48$, $p<.001$) and the pooled online student sample ($z=2.18$, $p=.029$). In TG, only the difference to the broader population sample is significant ($z=9.45$, $p<.001$).

A5 Replication: Trustees and Ultimatum Game

We can further substantiate our findings using second-mover behavior in TG. After receiving a monetary transfer (partly from a trustor and partly from the experimenter), a trustee can decide how much of the pie (0–100%) she passes back to the trustor. To elicit back-transfers in the TG, we employed the strategy method (Rauhut and Winter 2010), asking trustees to respond to all possible transfers. We refrained from using the standard measure of trustee behavior (second player’s response to actual transfer) in order not to censor the response variable.

In Studies 1 and 2, we further ran the Ultimatum Game (UG) as a third decision situation. We dropped UG in our non-student Studies 3 and 4 to sustain online participants’ attentiveness and to cap processing times to ranges typical for the respective panel or platform. In UG, a proposer receives a stake and can decide how much of the pie (0–100%) she offers to a responder. The responder can then accept (and both receive their share) or decline the offer (and both receive nothing). In contrast to DG, UG introduces fear of sanction into first movers’ decision situation. Proposers need to take into account responders’ option of rejecting offers that fall short of a “fair” split. Responders’ acceptance or rejection of offers, in turn, measures willingness to pay for the punishment of proposers’ unfair behavior. Again, we measured response behavior in UG using the strategy method: Here, we asked second movers to accept or reject each feasible offer, allowing us to determine responders’ minimal acceptable offer (MAO).

Figure A3 summarizes our findings for second movers in TG and for both roles in UG. Pooled across Studies 1–4, trustees return on average 43.2% of their received transfers. In his compendium on experimental games, Camerer (2003:86-88) reports trustees as typically returning about 100% of trustors’ unmultiplied investment. This, depending on the multiplication factor, amounts to 33–50% of responders’ actual received transfer. Pooled across Studies 1 and 2, mean offer in UG is 45.9% of the individual endowment. Responders’ MAO on average amounts to 27.4%. A meta-analysis of 75 UG-results reports mean offers of 40.4% (Oosterbeek, Sloof, and van de Kuilen 2004), with responders typically rejecting offers below 20% about half of the time (Camerer 2003:49).

Most important for our study is that the transportability of quantitative results fails for another laboratory measure of prosocial behavior: For trustees’ back-transfers, four out of seven pairwise comparisons reveal significant differences. On the other hand, for the decision situations faced by both UG proposers and responders, not a single pairwise comparison is significant.

In our two student pools (Studies 1 and 2), we find significantly different rates of trustees’ back-transfers ($t=2.60$, $p=.009$). In contrast, *pool generalizability* holds for the UG: Both proposers’ offers ($t=1.13$, $p=.258$) and responders’ MAOs ($t=0.11$, $p=.911$) are not significantly

Figure A3: Quantitative results from additional decision situations.

Note: Shaded bars show unconditional means of second-mover transfers in the Trust Game (TG) and first-mover and second-mover behavior in the Ultimatum Game (UG). Blank bars represent conditional means keeping underlying socio-demographics constant. We include 95% confidence intervals and seven pairwise comparisons (t -tests). *** $p < .001$, ** $p < .01$, * $p < .05$, n.s.=nonsignificant.

different between locations. The cross-location gap we were able to replicate for second movers in TG is absent for UG irrespective of whether we control for socio-demographic differences across students pools.

With respect to *sample generalizability*, trustees reproduce our earlier finding on the limited transportability of quantitative results: Back-transfers in the nationwide sample differ significantly from the student-online benchmark ($t=2.98$, $p=.003$). Similarly, comparing the student-online benchmark to participants at MTurk, we found *context generalizability* violated ($t=2.33$, $p=.020$). These results further call into question the transportability of quantitative results in social science lab research.

Corroborating our favorable result on *mode generalizability*, we find no measurement effects of lab vs. online data collection (Study 2). At both locations, data collected from trustees ($t=1.65$, $p=.100$ in Leipzig; $t=0.59$, $p=.557$ in Munich), proposers ($t=0.48$, $p=.630$ in Leipzig; $t=0.45$, $p=.653$ in Munich), and responders ($t=0.04$, $p=.969$ in Leipzig; $t=0.10$, $p=.918$ in Munich) do not differ, on average, between lab and online sessions.

A6 Instructions

Dictator Game Instructions

Task 1

Please read these instructions carefully.

For this decision task, you have been randomly matched to interact with another Mechanical Turk worker. None of you will know the other participant's worker ID.

We will provide \$2 to each pair of players.

Player A decides how to divide this money between him- or herself and Player B. Player A must allocate between \$0 and the total \$2 to Player B.

Player B takes home as a bonus whatever Player A allocates to him or her. Player A gets whatever he or she does not allocate to Player B.

You will see an example on the next page.

Next

Dictator Game Animation

Task 1

Here is an example:

Once Player A has made a decision, the interaction is over. Then, neither player will be able to affect the other's earnings.

You will be Player A. You will take three decisions in this task.

Each time, you will be matched to another Mechanical Turk worker. You will not know the other participants' worker IDs. The other participants will not know your worker ID.

Back

Next

Dictator Game Decision

Task 1, decision 1

You are **Player A**.

Player B is a Mechanical Turk worker **from the United States of America**.

Please divide the **\$2**.

The decision you take will determine the amount of money you earn from this HIT.

Which amount do you send to Player B?

\$0.00	\$0.20	\$0.40	\$0.60	\$0.80	\$1.00	\$1.20	\$1.40	\$1.60	\$1.80	\$2.00
☐	☐	☒	☐	☐	☐	☐	☐	☐	☐	☐

[Show instructions again.](#)

Next

Dictator Game Control Questions

Task 1

The following control questions are designed for your understanding.

You must answer these questions correctly to have the HIT approved.

In case of mistakes, you may correct your answers.

Suppose, Player A sends \$0.20 to Player B. In this case, what is the bonus Player B takes home?

\$0.00	\$0.20	\$0.40	\$0.60	\$0.80	\$1.00	\$1.20	\$1.40	\$1.60	\$1.80	\$2.00
☐	☒	☐	☐	☐	☐	☐	☐	☐	☐	☐

Suppose, Player A sends \$1.60 of his or her \$2.00 to Player B. In this case, what is bonus Player A takes home?

\$0.00	\$0.20	\$0.40	\$0.60	\$0.80	\$1.00	\$1.20	\$1.40	\$1.60	\$1.80	\$2.00
☐	☐	☐	☒	☐	☐	☐	☐	☐	☐	☐

✗ Wrong answer.
Please correct.

Trust Game Instructions

Task 2

Please read these instructions carefully.

For this decision task, you have been randomly matched to interact with another Mechanical Turk worker, different from the persons you interacted with before.
None of you will know the other participant's worker ID.

We will provide **\$1** to each player.

Player A decides how to divide his or her amount between him- or herself and Player B.

Player A must allocate between **\$0** and the total **\$1** to Player B.

This website then doubles the allocated amount and sends it to Player B.

Remember, Player B has (just like Player A) already received **\$1** from the experimenter.

In addition, Player B receives the doubled amount allocated to him or her by Player A.

Then, Player B decides how much of the doubled amount (between **0%** and **100%**) he or she sends back to Player A.

Player A takes home as a bonus **\$1** minus his or her allocation to Player B plus the amount Player B sends back.

Player B gets **\$1** plus the doubled amount allocated by Player A minus the amount he or she sends back to Player A.

You will see an example on the next page.

Next

Trust Game Animation

Task 2

Here is an example:

Once both Players have made their decisions, the interaction is over.

Then, neither player will be able to affect the other's earnings.

In this task, you will participate in two roles: First you will take three decisions as Player A.

Each time, you will be matched to another Mechanical Turk worker.

Then, paired with a new Mechanical Turk worker, as Player B.

You will not know the other participants' worker IDs. The other participants will not know your worker ID.

Back

Next

Trustor Decision

Task 2, decision 1

You are Player A.

Player B is a Mechanical Turk worker from the United States of America.

Please divide the \$1.

The decision you take will determine the amount of money you earn from this HIT.

Which amount do you send to Player B?

\$0.00	\$0.10	\$0.20	\$0.30	\$0.40	\$0.50	\$0.60	\$0.70	\$0.80	\$0.90	\$1.00
☐	☐	☐	☐	☐	☒	☐	☐	☐	☐	☐

[Show instructions again.](#)

Next

Trustee Decision

Task 2

You are now Player B.

You will be randomly matched to interact with another person, different from the persons you just interacted with.

While Player A divides his or her \$1, please decide on how much you are willing send back to Player A for each of the possible amounts.

You cannot change your decision once you will be informed about Player A's actual allocation.

	0	10	20	30	40	50	60	70	80	90	100	
	%	%	%	%	%	%	%	%	%	%	%	
Player A allocates \$0.00 to you. You receive \$0.00.												You earn \$1.00. In addition, you'd earn: \$0.00
Player A allocates \$0.10 to you. You receive \$0.20.	☐	☐	☒	☐	☐	☐	☐	☐	☐	☐	☐	\$0.16
Player A allocates \$0.20 to you. You receive \$0.40.	☐	☐	☐	☒	☐	☐	☐	☐	☐	☐	☐	\$0.28
Player A allocates \$0.30 to you. You receive \$0.60.	☐	☐	☐	☒	☐	☐	☐	☐	☐	☐	☐	\$0.42
Player A allocates \$0.40 to you. You receive \$0.80.	☐	☐	☐	☐	☒	☐	☐	☐	☐	☐	☐	\$0.48
Player A allocates \$0.50 to you. You receive \$1.00.	☐	☐	☐	☐	☐	☒	☐	☐	☐	☐	☐	\$0.50
Player A allocates \$0.60 to you. You receive \$1.20.	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
Player A allocates \$0.70 to you. You receive \$1.40.	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
Player A allocates \$0.80 to you. You receive \$1.60.	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
Player A allocates \$0.90 to you. You receive \$1.80.	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
Player A allocates \$1.00 to you. You receive \$2.00.	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	

[Show instructions again.](#)

Next

A7 Anonymization procedures in the two laboratories

Leipzig

Munich

Note: The top row shows the low-anonymity condition (no blinds, payoff at the experimenter's desk). The middle row shows the blinds used to create anonymity toward other participants (standard anonymity). The bottom row shows our payment setup to create anonymity toward the experimenter (high anonymity).

References

- Andreoni, James 1995. "Cooperation in Public-goods Experiments: Kindness or Confusion?" *American Economic Review* 85(4):891-904.
- Camerer, Colin F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. New York: Sage.
- Engel, Christoph 2011. "Dictator Games: A Meta Study." *Experimental Economics* 14(4):583-610.
- Etzioni, Amitai 2015. "The Moral Effects of Economic Teaching." *Sociological Forum* 30(1):228-33.
- Fréchette, Guillaume R. 2016. "Experimental Economics Across Subject Populations." Pp. 435-80 in *The Handbook of Experimental Economics, Vol. 2.*, edited by J. Kagel and A. Roth. Princeton, NJ: Princeton University Press.
- Johnson, Noel D. and Alexandra A. Mislin 2011. "Trust Games: A Meta-analysis." *Journal of Economic Psychology* 32(5):865-89.
- Levitt, Steven D. and John A. List 2007. "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?" *Journal of Economic Perspectives* 21(2):153-74.
- Oosterbeek, Hessel, Randolph Sloof, and Gijs van de Kuilen 2004. "Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-analysis." *Experimental Economics* 7(2):171-88.
- Rand, David G., Alexander Peysakhovich, Gordon T. Kraft-Todd, George E. Newman, Owen Wurzbacher, Martin A. Nowak, and Joshua D. Greene 2014. "Social Heuristics Shape Intuitive Cooperation." *Nature Communications* 5:3677.
- Rauhut, Heiko and Fabian Winter 2010. "A Sociological Perspective on Measuring Social Norms by Means of Strategy Method Experiments." *Social Science Research* 39(6):1181-94.
- Recalde, Maria P., Arno Riedl, and Lise Vesterlund 2018. "Error Prone Inference from Response Time: The Case of Intuitive Generosity in Public-good Games." *Journal of Public Economics* 160:132-47.
- Tinghög, Gustav, David Andersson, Caroline Bonn, Harald Böttiger, Camilla Josephson, Gustaf Lundgren, Daniel Västfjäll, Michael Kirchler, and Magnus Johannesson 2013. "Intuition and Cooperation Reconsidered." *Nature* 498:E1-E2.
- Wiepking, Pamala and René Bekkers 2012. "Who Gives? A Literature Review of Predictors of Charitable Giving. Part Two: Gender, Family Composition and Income." *Voluntary Sector Review* 3(2):217-45.