Judgment and Decision Making, Vol. 17, No. 2, March 2022, pp. 263–283

"When in Rome": Identifying social norms using coordination games

Erin L. Krupka* Roberto Weber† Rachel T. A. Croson‡

Hanna Hoover§

Abstract

Previous research in economics, social psychology, and sociology has produced compelling evidence that social norms influence behavior. In this paper we apply the Krupka and Weber (2013) norm elicitation procedure and present U.S. and non-U.S. born subjects with two scenarios for which tipping and punctuality norms are known to vary across countries. We elicit shared beliefs by having subjects match appropriateness ratings of different actions (such as arriving late or on time) to another randomly selected participant from the same university or to a participant who is born in the same country. We also elicit personal beliefs without the matching task. We test whether the responses from the coordination task can be interpreted as social norms by comparing responses from the coordination game with actual social norms (as identified using independent materials such as tipping guides for travelers). We compare responses elicited with the matching tasks to those elicited without the matching task to test whether the coordination device itself is essential for identifying social norms. We find that appropriateness ratings for different actions vary with the reference group in the matching task. Further, the ratings obtained from the matching task vary in a manner consistent with the actual social norms of that reference group. Thus, we find that shared beliefs correspond more closely to externally validated social norms compared to personal beliefs. Second, we highlight the importance that reference groups (for the coordination task) can play.

Keywords: norms, coordination games, experiment

*University of Michigan School of Information, 105 S. State Street, Ann Arbor, MI 48109. Email: ekrupka@umich.edu. https://orcid.org/0000-0002-9194-7501.

†University of Zurich Department of Economics, Blümlisalpstrasse 10 8006 Zürich, Switzerland. Email: roberto.weber@econ.uzh.ch. https://orcid.org/0000-0001-8133-8131.

‡Executive Vice President and Provost, University of Minnesota, 234 Morrill Hall, 100 Church Street Minneapolis, MN 55455. Email: provost@umn.edu. https://orcid.org/0000-0002-3555-3970.

§University of Michigan School of Information, 105 S. State Street, Ann Arbor, MI 48109. Email: hooverha@umich.edu. https://orcid.org/0000-0002-7919-7867.

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press

1 Introduction

Social norms, defined as collective perceptions of appropriate behavior, are an important topic of study within psychology, sociology, and economics. Within economics, social norms offer a possible explanation for behavior that otherwise appears inconsistent with self-interest and also help explain human coordination when multiple possible equilibria exist (Sugden, 1995; Binmore & Samuelson, 2006; Gintis, 2014).

As social norms have gained prominence as a topic of study in their own right or as an explanatory factor for observed behavior, recent work in economics specifically targets the need for new tools that empirically measure social norms. Krupka and Weber (2013) develop an incentive compatible norm elicitation protocol that avoids the potential hypothetical bias of unincentivized techniques and captures a key feature of social norms: that they are jointly recognized (Harrison & Rutström, 2008; Roth & Kagel, 2016).[1] [2] The protocol asks subjects to play an incentivized coordination game that rewards subjects for coordinating on the same rating of how appropriate an action is in a given choice context. This method assumes that collectively recognized social norms create focal points in the matching game — if there is a social norm that some actions are more or less socially appropriate, respondents attempting to match others' appropriateness ratings are likely to rely on this shared perception to help them do so. Thus, the incentives in the coordination game elicit collective perceptions of appropriateness which they interpret as an empirical measure of the social norm.

The primary research question we address is whether the responses from the coordination task are appropriately interpreted as empirical measures of social norms.[3] We test this

[1]For example, Reuben and Riedl (2013) used an unincentivized questionnaire that asks respondents to put themselves in the position of an "uninvolved neutral arbitrator". The questionnaire asked respondents about their normative beliefs regarding the "fair" amount that a group member should contribute to a public good described in the questionnaire (in a second question, they also elicited the fair amount condition on another's contribution). These two papers join a stream of literature in psychology that also uses unincentivized questions to elicit norms (e.g., Opp, 2001; Cialdini et al., 1990).

[2]Schram and Charness (2015) used advice by a payoff-independent third party as a mechanism for (a) creating a collective perception of the social norm among advisor and advisee and (b) identifying the decision rule that constitutes the social norm for the particular situation over which advice is rendered. Bicchieri and Chavez (2010) and Bicchieri and Dimant (2019) used a different method to identify social norms. In this method, a first group of subjects reads a vignette and then reports what they think one "should do" (i.e., their personal normative beliefs). Then a second group of subjects predicts what the majority of subjects in the first group indicated that one "should do". This second elicitation is incentivized and subjects are paid a monetary reward if their normative expectations match the personal normative beliefs of the first group of subjects. This method uses the distribution of beliefs (rendered by the second set of subjects) about the personal norms in the reference group to proxy for the social norm. However, one might reasonably conclude that this approach measures the jointly held beliefs (of the second set of subjects) about the distribution of personal beliefs rather than the jointly held beliefs about the distribution of jointly held beliefs (something the Krupka & Weber, 2013, approach measures).

[3]Using coordination games to elicit injunctive norms, Burks and Krupka (2012) demonstrated that the responses obtained from the matching task are relatively insensitive to variations in three other likely alternative focal points for coordination: a subject's beliefs about what others actually do, beliefs about what they themselves would do, or observing the choices of several other subjects.

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press

in three ways. First, we test whether responses using the coordination mechanism are systematically different from responses when no such coordination mechanism is in place; thus, we can test whether the coordination game yields different responses from a subject than the unincentivized "what do you personally believe" survey approach. Second, we test whether foreign-born nationals coordinating with foreign-born nationals respond differently in the coordination task than U.S. born nationals coordinating with U.S. born nationals; because foreign-born nationals have different norms, the responses of the two groups should differ in a fashion consistent with the difference in norms. Finally, we test whether responses from the coordination task track real-world social norms that we identify using independent source material (for, e.g., traveler's guides on country tipping norms). We will refer to these as ex-ante identified norms in what follows.

We find that appropriateness ratings for different actions vary with the reference group in the matching task. Further, the ratings obtained from the matching task vary in a manner consistent with the ex-ante identified social norms of that reference group. Thus, we find that shared beliefs correspond more closely to externally validated social norms than to personal beliefs.

Our findings also contribute towards the measurement of social norms by highlighting the central role of the reference group in determining the salient social norm. We examine two situations where an individual may belong to multiple social groups for which different social norms exist for the same situation. The experiment exploits the natural variation of social norms between two salient reference groups for participants – their fellow students at the university and the subset of students born in the same country. The coordination game we use to elicit norms is well suited to this task as it has the advantage of allowing us to systematically vary the reference group the respondent uses in their matching tasks.

Our findings join a recent study that elicits social norms for different reference groups and distinguishes social norms from personal norms. Burks and Krupka (2012) show that social norms, elicited using the coordination game, are distinct from personal opinions and they demonstrate the separate effect of personal opinions and social norms on behavior. Unlike in Burks and Krupka (2012), this paper uses two well-articulated real-world social norms — tipping and punctuality — to benchmark responses in the coordination game against externally validated norms. We ask about tipping and punctuality norms because it is possible to use independent source material to obtain ex-ante descriptions of the social norm for these domains and because the norms for these domains vary by country (Lynn and Lynn (2004); Lynn (2006)). Further, in this paper we show an additional way in which we can adapt the elicitation protocol to vary reference groups to study social norms in groups where culture plays a significant role in shaping those norms.

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press

2 Definition of social and personal norms

A long tradition in psychology distinguishes several different kinds of norm constructs: social norms that describe what one ought to do, descriptive norms that describe what is regularly done and personal norms (Deutsch & Gerard (1955); Schwartz (1973); Cialdini et al. (1991)). We follow this literature in distinguishing what one "ought" to do, or injunctive norms, from customs or actions that people regularly take, or descriptive norms (Bicchieri (2005); Deutsch and Gerard, (1955)). Both kinds of norms influence behavior (Herrbach & Mignonac, 2007; Cialdini et al., 1990; Krupka & Weber, 2009; Bicchieri & Xiao, 2009); Gneezy et al., 2010). However, our primary focus in this paper is on injunctive norms, i.e., those described by Elster (1989) as prescribing what one "should do" or "should not do", and on personal norms. We define injunctive norms as a socially shared understanding regarding the in/appropriateness of a particular behavior.

This definition emphasizes that social norms are characterized by a "shared" understanding (See, for example, Elster, 1989; Bicchieri, 2005; Miller and Prentice, 1996; Young, 2007).[4] For example, Bettenhausen and Murnighan (1991, p. 21) note that norms are "socially shared guidelines to accepted and expected behavior." Fehr and Gaechter (2000, p. 166) write that norms are "based on a socially shared belief about how one ought to behave."

Distinguishing between collectively held norms and personal normative beliefs (personal beliefs) is equally important, both in practice and in testing theoretical predictions. *Social* norms are characterized by shared knowledge while personal beliefs need not be known to any other member of the group and are expectations one holds of oneself (Schwartz, 1973; Elster, 1989). Schwartz (1977) describes personal norms as self-expectations for behavior constructed in specific situations on the basis of generalized internalized values (however, personal norms are similarly characterized in Elster, 1989, and Posner & Rasmusen 1999).[5] It follows that an individual's personal normative beliefs need not track a group norm

[4]Bernheim (1994) states that "in practice, for any given society, one may observe many cohesive subgroups, each with its own distinct norm". Miller and Prentice (1996) state that "a social norm is *an attribute of a group* that is considered to be both descriptive of and prescriptive for its members [Emphasis added]." Bicchieri (2005) writes "A behavioral rule R can be a social norm for one population P and not for another population P'." See also Jones (1984), Sugden (2000), Akerlof and Kranton (2005), Goette et al. (2006), Charness et al. (2007).

[5]Schwartz (1968) and Schwartz (1977) defines personal norms as "self-expectations that are based on internalized values" that are experienced as feelings of personal obligation to engage in a certain behavior. Upon distinguishing between personal (or "private") norms and social norms, Elster writes that "private norms . . . are not shared with others." In the literature that followed, "moral" and "personal" norms are both used interchangeably and Schwartz (1977) notes that it is unclear which is best. In the present study, we have opted to use the term "personal". In economics, there is also a tradition of distinguishing personal from social norms; this literature defines the differences through the nature of beliefs that support each. Personal norms measure an actor's unconditioned (on other's normative expectations) normative convictions. Social norms are characterized by an actor's beliefs about what others expect of them and what they expect of others (Catola et al., 2021, Bicchieri et al. 2022). Thus, for example, Bicchieri differentiates social from personal norms (or "personal opinions") by the fact that social norms "have no reality other than our beliefs that others behave according to them and expect us to behave according to them."

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press

(Bicchieri, 2005; Young, 2007). Thus, as an example, when a new student from another country joins a U.S. university he may recognize that the social norm for tipping in the U.S. is between 12% and 15% for lunch, but may personally believe that tipping 5% after a restaurant lunch meal is appropriate. Therefore, if one were to ask such an individual: "Do you think that tipping 5% is something that you ought to do?" they might respond on the basis of what they personally agree with (5%), even if they know that the relevant social norm in the U.S. is higher (15%).

The norm elicitation method first proposed by Krupka and Weber (2013) captures this joint recognition by using a specific behavioral economic experiment — a coordination game over rating the appropriateness of actions for a scenario that they describe to subjects. Camerer and Fehr (2004, p. 34) note that coordination games can be paired with economic incentives to reveal shared understanding and suggest that experimental paradigms, such as simple coordination games, can prove useful for measuring dimensions of shared perception.[6] From a game-theoretic point of view, matching games such as the one used in the norm elicitation protocol, have a number of equilibria and nothing intrinsic to the game makes one equilibrium favored (or focal) over the other. Schelling (1980) theorized and Mehta et al. (1994) and Sugden (1995) demonstrated that prominence derived from common culture and shared experiences can create focal points. Krupka and Weber (2013) showed that in their games, norms created focal points that subjects used to successfully coordinate. Further, it follows that personal norms, captured without an incentive and without the coordination game, may significantly vary from the views that are understood by group members to constitute the collective norm.

3 Experimental design and hypotheses

Table 1 describes the experimental design. In this study, respondents participate in six different stages or tasks.

Each participant went through all six stages. Participants always read about the tipping vignette first and were either presented with a coordination game in which the target was to match with other participants from the same university or with others who were born in the same country. The order of the targets for matching was randomized. Once they completed both coordination games, they were asked to rate each action according to what they personally believed. The coordination tasks were incentivized such that they received payment for correctly matching with the target and personal beliefs were not. This pattern of tasks was repeated for the punctuality vignette. The study also included a demographics survey.

[6]Crawford et al. (2008) demonstrate how the power of focal points is considerably weakened by payoff asymmetries. In our elicitation experiment, we use symmetric payoffs.

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press

Table 1: Experimental design overview.

Stage for each participant	Vignette	Target for matching ratings	Order Counter-balanced	Incentives to match responses?
1.	Tipping	Coordinate ratings with other participants at the university	Randomized the order of which target participants matched with	Yes
2.		Coordinate ratings with other participants born in the same country		Yes
3.		No target (aka. personal beliefs)	Always after matching	No
4.	Punctuality	Coordinate ratings with other participants at the university	Randomized the order of which target participants matched with	Yes
5.		Coordinate ratings with other participants born in the same country		Yes
6.		No target (aka. personal beliefs)	Always after matching	No

The tipping vignette is always the first vignette subjects encounter as they accomplish tasks in stage 1–3 in Table 1.[7] In the "tipping vignette", subjects read a scenario where a third party must decide what percentage of a $10 lunch bill to tip. Subjects read the following vignette: "Individual A has just finished lunch at a restaurant. The bill for lunch is $10.00. Individual A must decide how much, if any, tip to leave. Individual A has seven possible choices which are indicated in the table below. Individual A can choose only one of these options." Subjects are told to think about the following possible actions for Individual A: give no tip (A pays $10 and a $0 tip), give a 5% tip (A pays $10 and a $0.50 tip), give a 7% tip (A pays $10 and a $0.70 tip) give a 10% tip (A pays $10 and a $1 tip), give 12% tip (A pays $10 and a $1.20 tip), give 15% tip (A pays $10 and a $1.50) tip, give 20% tip (A pays $10 and a $2.00 tip).

The punctuality vignette is always the second vignette subjects encounter as they accomplish tasks in stage 4–6 in Table 1. In the "punctuality vignette" subjects read about a third party who must decide on whether to arrive early, on time, or late to a meeting at the library. Subjects read the following vignette: "Individuals A and B have agreed to meet in the library to work on an assignment. They have agreed to meet at a particular time. Individual A has to decide at what time to arrive for the meeting. Individual A has seven possible choices which are indicated in the table below. Individual A can choose only one

[7]The full experimental instructions are at http://ekrupka.people.si.umich.edu/.

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press

of these options." Subjects were asked to think about the following possible actions for Individual A: arrive 10 minutes early, 5 minutes early, exactly on time, 5 minutes late, 10 minutes late, 20 minutes late or 30 minutes late.

In each case, tipping or punctuality, if they are playing the coordination game, then they must match ratings with another participant who is from their university or born in the same country as they were to receive a bonus payment (stages 1–2 and 4–5 in Table 1). The coordination task asks them to rate the possible actions according to how socially appropriate they are. Respondents judge the social appropriateness of *each* action in the vignette on a four-point scale that ranges over "very socially inappropriate", "somewhat socially inappropriate", "somewhat socially appropriate" to "very socially appropriate."[8,9] They do the coordination ratings two times. If they are matching with another randomly selected university student, then they are paid if they successfully select the same appropriateness rating as that participant. If they are matching ratings with another participant born in the same country, then they are paid if they successfully select the same appropriateness rating as that participant. In both coordination tasks, respondents play a "pure matching" coordination game (Schelling, 1980; Mehta et al., 1994) in which their goal is to anticipate the extent to which others in the reference group will rate an action as socially appropriate or inappropriate, and to respond accordingly.

Table 2 provides an example of a table subjects used to record their responses when they were matching their responses with another randomly selected university subject participating in the experiment. The 'x' marks in the table are for exposition purposes and illustrate how a participant could have responded.

After playing the two coordination tasks, the third task for each vignette is not a coordination game but simply asks them to state their personal beliefs about how appropriate the actions are. There is no reference group in this stage (3 and 6, in Table 1) and it always comes after the two coordination tasks have been completed for the vignette.

As can be seen in Table 1, the order of the vignettes was not counterbalanced; tipping always came first. However, the order of the coordination games was counterbalanced such that some were asked to first match with another university student and then with another student born in the same country, while for others this order was reversed. After the two coordination games, subjects provided their own appropriateness ratings without trying to match anyone else's response (they had no target and were paid no incentive). This three-stage sequence was completed twice in the experimental session, once for the

[8]The decision to have only four appropriateness categories was made after considering the tradeoff between having too few (in which case it would be harder to discriminate between degrees of appropriateness) and having too many (in which case it might be too difficult for subjects to match on the social norm, perhaps leading them to attempt to match using other focal principles). Further, we omitted the "neutral" category, as this would have been a focal point separate from the focal point stemming from the social norm.

[9]To provide a clear objective for subjects, we give the following description of socially appropriate: "By socially appropriate, we mean behavior that most people agree is the 'correct' or 'ethical' thing to do. Another way to think about what we mean is that if Individual A were to select a socially inappropriate choice, then someone else might be angry at Individual A for doing so."

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press

TABLE 2: Example of a response reporting form used in the coordination game tor the tipping vignette. This table gives an example of the response form and response format that was used to elicit participants' beliefs about, in this example, tipping. For example, this hypothetical respondent has indicated that they believe that most others in the target group would say that giving "no tip" would be "very socially inappropriate", while most others in the target group would say that giving "20% tip" would be "very socially appropriate".

Individual A's choice	*Very socially Inappropriate*	*Somewhat socially inappropriate*	*Somewhat socially appropriate*	*Very socially appropriate*
Give no tip	X			
Give a 5% tip	X			
Give a 7% tip		X		
Give a 10% tip			X	
Give a 12% tip			X	
Give a 15% tip				X
Give a 20% tip				X

tipping vignette and once for the punctuality vignette, such that each subject completed six segments in total and a seventh segment asked them some basic demographic information.

Payment in the coordination game occurred a week later and consisted of a $10 participation fee for all participants and the possibility of earning additional money if their response was selected for payment in the coordination tasks. To calculate payments for the coordination task, the experimenter randomly selected one participant and matched their responses to the responses provided by another subject from each of the two reference groups. Specifically, the chosen subject's ratings in the coordination games would be matched with another randomly selected subject from their university and then with a second randomly selected subject who was born in the same country as the chosen subject.[10] The selected participant would receive an additional $1 for every rating that was the same rating as the matched responses. As there are 7 response categories and 4 coordination games, the chosen subject had the opportunity to make an additional $28 if all the responses coincided with the two matched persons' responses. In addition to the coordination incentive, the selected subject would also receive $20 if every question was completed. This payment scheme provides a salient incentive for participants to accurately consider how other participants will judge the appropriateness of any given action.

We interpret subject responses in the following manner. When subjects are coordinating with a randomly selected university student participating in the experiment, then this technique elicits the subject's belief about the normative evaluations of another student at

[10]Subjects were also told that if a match could not be made on the country level, we would select a subject born in the same region.

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press

the university, and in aggregate is a proxy measure for the norm for the university. If the subject's reference group is from the set of participants who were born in the same country, then this technique elicits the subject's belief about the normative evaluations of his nationals, and in aggregate is a proxy measure for the norm for nationals from that country. If there is no incentive structure and subjects are asked to state their personal beliefs, then this technique elicits the subject's personal belief and, in aggregate, identifies a distribution of personal beliefs for a group.

3.1 Hypotheses

Our coordination game treatments consist of subjects matching their responses with (1) a subject selected randomly from the sub-set of subjects who were born in the same country and (2) a randomly selected subject from all university students participating. Because our university is located in the U.S., the university norm is likely to map very closely to a U.S. norm (both because most students will be from the US and because of the geographic location of the university). We tested the following hypotheses:

H1: [Personal beliefs will differ from ratings in the coordination treatments] Personal beliefs for foreign-born respondents will differ significantly from the ratings they provide in the coordination task where the target is another university student.

H2: [Ratings(F,F) will differ from Ratings(US,US)] For either tipping or punctuality, the ratings of foreign-born nationals coordinating with foreign-born nationals will differ from the ratings of U.S. born nationals coordinating with U.S. born nationals.

H3: [Ex-ante identified norms will predict the direction of ratings changes] Respondent's ratings when coordinating with foreign-born (ratings(F,F)) will be different from ratings when U.S. born nationals coordinate with other U.S. born nationals (ratings(US,US)) in a manner that is consistent with ex-ante identified differences in the relevant norms.

3.2 Procedure

Students from a northeastern business school in the U.S. were recruited to take part in our experiment for partial fulfillment of their class requirements and were paid a flat show-up fee of $10. A total of 165 participants were recruited. Sessions were conducted using groups of 17 to 35 participants and included another unrelated decision task that always took place before this experiment. Instructions were read aloud and participants were informed of the incentive structure for the task. Before beginning the experiment, subjects read an example scenario which demonstrated how to fill out the tables and how the matching would be executed. One week after all the sessions had been conducted the participant who was selected to receive additional payment was notified publicly in class and paid.

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press

4 Results

The sample size for our analysis was 155 subjects who were born in 19 different countries.[11] Table 3 describes the distribution of cultural-geographic clusters of our sample based on country of birth and ex-ante social norms identified through reputable sources. Most of our subject pool was born in the USA (70.3%). The largest non-U.S. nationality represented was China (9.6%) followed by Singapore (3.2%) and India (2.5%). Columns 5 and 6 give an overview of each country's tipping and punctuality norms.

Each country's tipping norm was identified using at least three different sources. We relied heavily on Fodor's "How to Tip" (2002) travel guide and Magellan's World Wide Tipping Guide as well as Nancy Star's book, The International Guide To Tipping, on tipping in the service industry. Star obtained her tipping guidelines by asking the service providers to fill out a questionnaire in which, among other things, they were asked to state their 'average' tip. Despite its print date (1988), the tipping ranges she reports are consistent with those stated in our other sources.

Point estimates of punctuality norms were more difficult to obtain but it was not difficult to learn whether a country generally viewed punctuality favorably or unfavorably. Levine et al. (1980) asked 107 U.S. students "at what time the average American would arrive" to "a lunch with a friend." Both U.S. males and females reported that the average American would arrive 2–2.5 minutes late. We consider this (arriving 2.5 minutes late) the punctuality norm for the U.S. While some studies report punctuality norms (Levine et al. 1980; Levine, 2008; White et al., 2011; Gelfand et al., 2011), we were unable to locate multiple references for point estimates.[12] For this reason we used a binary indicator that coded for whether a country was less punctual or more punctual than the U.S. Other country-specific punctuality norms were described as being later or earlier than the U.S. (where earlier typically means arriving on time, and later typically means arriving more than 5 minutes late). These norms were obtained from multiple references and internet resources dedicated to doing international business. There were no disagreements regarding how to classify a country's social norm relative to the U.S. from these sources.[13] Thus, we can interpret Table 3 as a table that describes the country-specific ex-ante identified social norms

To test H1, that personal beliefs differ from ratings in the coordination treatment, we construct a measure of the respondent's preferred action (denoted as pTip and pPunctual). This is a measure of the average tip percentage or minutes early/late that received the

[11]We collected data from 165 subjects, but 5 subjects were dropped because some demographic or identifying variable vital to the analysis was not entered. Five additional subjects failed to complete the entire survey.

[12]Results for countries that are less punctual than the U.S. are not sensitive to whether we code the U.S. as having a "zero minutes late" (i.e., "on time") norm or a "two minutes late" norm.

[13]Levine's research, as well as Basu and Jörgen's (2003) paper, were valuable resources for obtaining punctuality norms. We also consulted Morrison and Conaway (1994) guide on cross-cultural protocol regarding punctuality.

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press

TABLE 3: Descriptive statistics for the sample. This table displays the important descriptive characteristics of our data. In the final two columns we also note the ex-ante identified norm for the country. The ex-ante identified norms were collected from multiple independent sources.

				Ex-ante identified norms	
Country born	N	Percent	Geographic cluster	Tip as % of Bill	Punctuality (vs. U.S.)
USA	109	70.32	North America (n=112)	12%	2 minutes late
Canada	3	1.94		12%	+
Brazil	3	1.94	Central/South America (n=7)	9%	-
Peru	2	1.29		7%	-
Columbia	1	0.65		7%	-
Ecuador	1	0.65		10%	-
Russia	3	1.94	Europe (n= 6)	9%	-
Germany	1	0.65		2%	+
Turkey	1	0.65		9%	-
France	1	0.65		5%	+
Singapore	5	3.23	Asia (n= 21)	0%	+
China	15	9.68		0%	+
Taiwan	1	0.65		2%	+
India	4	2.58	South Asia (n=4)	7%	-
Kenya	1	0.65	Africa (n=3)	2%	-
Uganda	1	0.65		4.5%	-
South Africa	1	0.65		10%	-
U.A.E	1	0.65	Middle East (n=1)	0%	-
Cayman Isl.	1	0.65	Caribbean (n=1)	12%	-
Total	155	100			

maximum appropriateness rating in the personal belief rating task.[14] Furthermore, we let uTip and uPunctual be a participant's guess about the most appropriate action in the university-matching coordination task, and nTip and nPunctual be the participant's guess about the most appropriate action in the nationality-matching coordination task. For the

[14]Thus, for example, if a subject indicated that tipping 15% was personally very appropriate and that tipping 20% was personally very appropriate, then their personally preferred tip percentage, *pTip*, would be 17.5%.

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press

punctuality scenario, arriving late is coded with a negative sign (i.e., arriving 10 minutes late is coded as –10) and arriving earlier is coded with a positive sign (i.e., arriving 10 minutes early is coded as 10).

We start by focusing on non-U.S. born subjects because these participants likely have personal beliefs that differ from those held at a university located in the U.S. When these subjects coordinate with university subjects, their guess about the most appropriate action should change to identify the university social norm. To test this, we compute the difference in ratings between pTip and uTip for each non-U.S. born subject and then test whether the mean of the distribution of these differences is zero.

Table 4 reports the means for pTip and uTip along with the mean of the distribution of difference between pTip and uTip (reported in the column labeled Avg(pTip–uTip)). These values are reported in columns 1 through 3.

Result #1a: Non-U.S. born subjects personally prefer tipping a lower amount than what they guess to be the most appropriate tipping amount when matching with university students. This is consistent with what we would predict from the norms identified in Table 1; U.S. tipping standards are high relative to most other countries.

Support for result 1a comes from Table 4 column 3. The mean difference (i.e., Avg(pTip–uTip)) is –0.991 and is significantly different from zero (t(45)=-3.170, p=0.003).

TABLE 4: Means Of Most Appropriate University Action And Personally Preferred Action In The Tipping Vignette. Columns 1 and 2 report the means of pTip and uTip and column 3 reports the mean of the distribution of differences between pTip and uTip. Standard errors are reported underneath the means in parentheses.

	pTip	uTip	Avg(pTip–uTip)
Non-U.S. Born (n=46)	15.570	16.560	–0.991
	(0.391)	(0.321)	(0.312)
U.S. Born (n=109)	16.940	16.660	0.275
	(0.278)	(0.257)	(0.239)

Subjects who were born within the U.S. have personal beliefs that align with their guess for the appropriate action when matching with university students. In this subsample, the mean of the distribution of differences between pTip and uTip is 0.275 (column 3). It is higher than when they coordinate with other university students but not significantly different (t(108)=1.150, p=0.253). In other words, U.S.-born subjects' personal belief about the most appropriate action is not different from their guess about the university norms for tipping.

Table 5 reports the same measures as Table 4, but for the punctuality scenario. We conducted paired t-tests to assess whether respondents' personally preferred punctuality

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press

differed from their best guess about the university's punctuality norm. However, because our punctuality norms are more crude (such that we know if a country's norms are to arrive earlier than the U.S. norm or to arrive later), we partitioned subjects into two groups: those who were born in countries where the ex-ante norm is to arrive later than the U.S. norm and those who were born in countries where the ex-ante norm is to arrive earlier than the U.S. norm (shown in Table 3).

TABLE 5: Means of most appropriate university action and personally preferred action in the punctuality vignette. Columns 1 and 2 report the means of pPunctual and uPunctual and column 3 reports the mean difference between pPunctual and uPunctual. Standard errors are reported in parentheses underneath the means.

	pPunctual	uPunctual	(pPunctual – uPunctual)
Arrive 'Early' (n=26)	4.230	3.850	0.385
	(0.515)	(0.421)	(0.359)
Arrive 'Late' (n=20)	3.430	4.750	–1.312
	(1.021)	(0.541)	(0.991)

Result #1b: Although Non-US born subjects' personally prefer arriving at a different time than what they guess to be the most appropriate arrival time when coordinating with university students, the means of the distributions of these differences are not significantly different from zero.

Support for result 1b comes from Table 5. For subjects born in 'later' countries, the mean difference between pPunctual and uPunctual is –1.313 and is not significantly different from zero (t(19)=–1.324, p=0.201).[15] Despite the lack of statistical significance, the direction of the difference indicates that these subjects prefer to arrive later in comparison to the guess of the most appropriate time of arrival when matching with university students. For subjects born in countries whose ex-ante norm is to arrive 'earlier' in comparison to the U.S. norm, the mean difference between their pPunctual and uPunctual responses is 0.385 and is also not significantly different from zero (t(25)=1.072, p=0.294).[16] Again, despite the lack of statistical significance, the direction of the difference indicates that these subjects prefer to arrive earlier in comparison to the guess of the most appropriate time of arrival when matching with university students.

To test H2, that the ratings of foreign-born nationals coordinating with foreign-born nationals will differ from the ratings of U.S. born nationals coordinating with U.S. born nationals, we begin with a visualization. The responses used in this part of the analysis are from subjects when they play the incentivized coordination game matching ratings with a

[15] A one sided t-test of (pPunctual – uPunctual)<0 is also n.s. (p=0.101).
[16] A one sided t-test of (pPunctual–uPunctual)>0 is also n.s. (p=0.147).

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press

reference group of other students who were born in the same country. We expect subjects to coordinate responses in such a way that identifies ex-ante norms corresponding to their relevant reference-group. For those born in the U.S., the ex-ante identified tipping norm is around 12%, whereas those not born in the U.S., Canada, or the Cayman Islands, the ex-ante norm is less than 12%. To visualize how appropriateness ratings differ amongst U.S.-born and non-U.S. born subjects, we plot the average appropriateness ratings by action separately by place of birth in Figure 1.

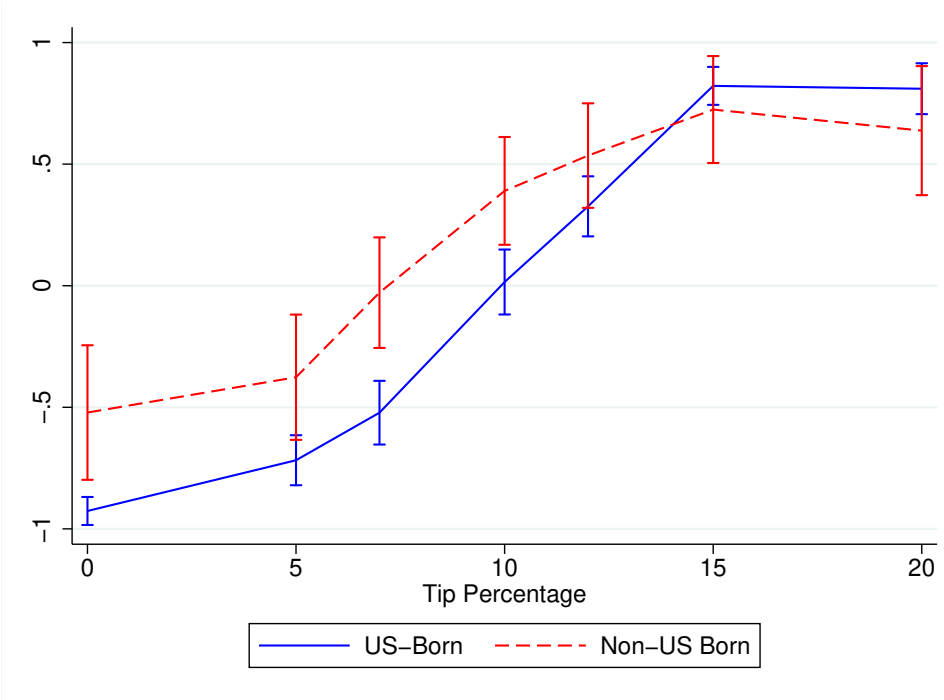

FIGURE 1: Average Rating of Tip Percentage by Respondent Country of Birth. This figure displays the average appropriateness rating by action and origin of birth for the tipping vignette during the same country born elicitation task. The y-axis reports the possible appropriateness ratings (ranging from very inappropriate, –1, to very appropriate, 1). The x-axis reports the tip percentages that were rated by participants. Confidence intervals are calculated as $\bar{x} \pm t_{n-1,\alpha/2} \cdot s/\sqrt{n}$ where $\bar{x}$ is the arithmetic mean, $t_{n-1,\alpha/2}$ is the $\alpha/2$ lower quantile of a t_{n-1} distribution. The sample size is n=46 for non-U.S. born and 109 for U.S.-born. The confidence level is α=5%. The sample standard deviation is denoted with the symbol s in the confidence interval calculation.

For tipping percentages below 10%, foreign-born subjects report higher appropriateness ratings for each action compared to U.S.-born subjects. In other words, U.S.-matching-U.S. subjects believed it was more inappropriate to leave a tip below 10%. For tips at or above 10%, the appropriateness ratings were not statistically different from one another.

For the punctuality scenario, we plot the average appropriateness ratings by each action and by country of birth in which the norm is to arrive 'on-time' and countries in which the

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press

norm is to arrive 'late' in Figure 4.[17]

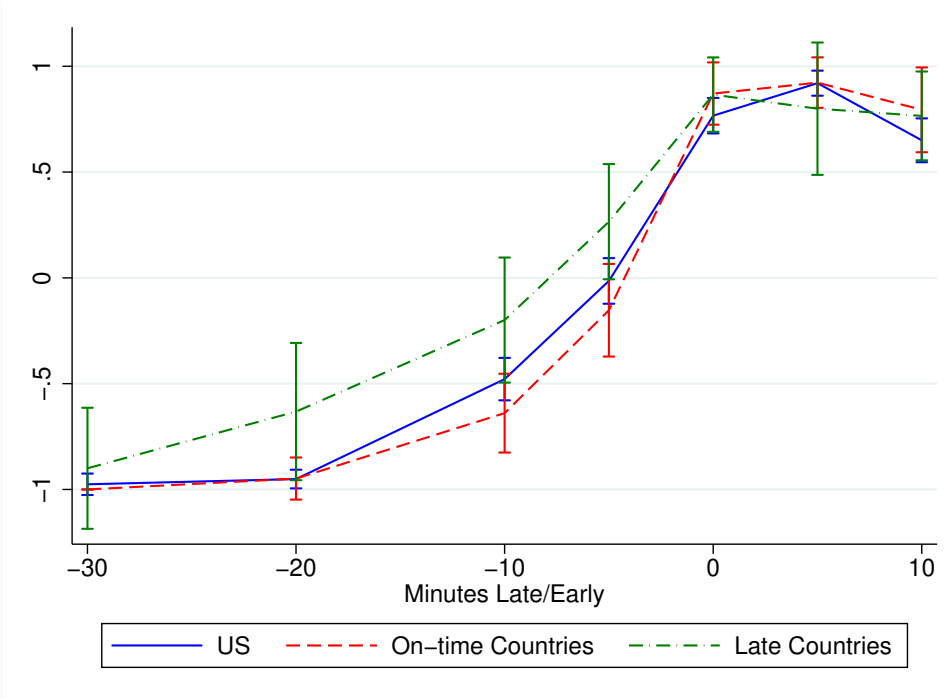

FIGURE 2: Average Ratings by Arrival Time by Country-Born and Punctuality Norm. This figure displays the average appropriateness rating by action in the punctuality vignette during the same country born elicitation task. The average ratings are broken out by whether the rater comes from the U.S., a late or an on-time country. The y-axis reports the possible appropriateness ratings (ranging from very inappropriate, –1, to very appropriate, 1). The x-axis reports different arrival times relative to a pre-agreed arrival time. Early arrival times include 30 minutes prior to (-30) etc. and late arrival times are 10 minutes after (10 on the x-axis). The plots are partitioned by responses from raters whose country of birth has an 'On-time' or 'late' norm or if it is the U.S. norm. The 'on-time' countries include Canada, Germany, France, Singapore, China, and Taiwan. 'Late' countries include Brazil, Peru, Columbia, Ecuador, Russia, India, Kenya, Uganda, South Africa, United Arab Emirates, and Cayman Islands. Confidence intervals are calculated as in Figure 1.

Figure 2 shows that most of the punctuality actions have indistinguishable appropriateness ratings for U.S.-born and 'on time' foreign-born subjects. When comparing U.S.-born subjects to 'late' countries, however, foreign-born subjects believe it is more appropriate to arrive late compared to U.S.-born subjects. The visual evidence suggests support for H2 (foreign-foreign pairs vs. U.S.-U.S.), and even some evidence for H3 (ex-ante norms predict direction of difference).

[17]('On-time' foreign countries include Canada, Germany, France, Singapore, China, and Taiwan while `'late' countries include Brazil, Peru, Columbia, Ecuador, Russia, India, Kenya, Uganda, South Africa, United Arab Emirates, and Cayman Islands.

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press

Result #2: Foreign-born subjects matching with foreign-born subjects reported different appropriateness ratings than U.S.-born subjects matching with U.S.-born subjects during the nationality-matching task (H2).

The mean of nTip reported by non-U.S. born subjects is 2.928 less than that reported by U.S. born subjects, which is statistically significant (independent samples t(153)=-4.857, p=0.000). This supports the inference that foreign-born subjects' best guess about their country tip norms is substantially lower than the guess of U.S. born subjects' guessing U.S. tip norms.

For the punctuality scenario, we broke the analysis into two sub-groups, those whose actual norm is to arrive later than the U.S. norm and those whose actual norm is to arrive earlier. Comparing 'later' non-U.S. subjects' nPunctual responses to U.S. subjects' nPunctual responses, we find the difference in means is –1.784 and is almost statistically significant at the 5% level (t(127)=–1.899, p=0.059). A one-sided two sample t-test may be more appropriate, however, given the directionality of the hypothesis ('late' countries view arriving late as more appropriate), and the one-sided test is significant (p=0.029). Comparing 'earlier' non-U.S. born subjects' nPunctual responses to U.S. born subjects' nPunctual responses, the difference in means is 0.360 and is not statistically different from zero (t(133)=0.577, p=0.565). The direction of the mean difference, however, is consistent with the differences in the actual norms.

Returning to Figures 1 and 2, we also have visual evidence supportive of H3 (that ex-ante identified norms will predict the direction of ratings changes). A simple way to test H3 is to test whether foreign-born subjects are responsive to their nation's norms when told that they will be matched with other nationals. Looking at responses of those matching with others born in the same country, we identify the tipping or punctuality norm for the nationality (using the information provided in Table 1) and the participant's best guess of their country's norms in the matching task with same nationality subjects (nTip or nPunctual).

Result #3: Respondent's ratings when coordinating with foreign-born ($ratings_{(F,F)}$) differ from ratings when U.S. born nationals coordinate with other U.S. born nationals ($ratings_{(US,US)}$) in a manner that is consistent with ex-ante identified differences in the relevant norms.

Our first form of support for result 3 comes from testing whether the correlation between a participant's guess about their own country's norm (when coordinating with a fellow national) is correlated with the ex-ante identified norm for their home country (i.e., with the actual norm of their home country). For the tipping scenario, the correlation between the best guess about their country's norm, nTip, and the actual norm is 0.402 (p=0.000). For the punctuality scenario, the correlation between the respondents' best guess about their country's norm, nPunctual, and the actual norm is 0.149 (p=0.064). These correlations show that subjects' coordination game responses, when coordinating with a fellow national, are consistent with the ex-ante identified country norms.

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press

We can also test whether foreign-born respondents adjust their guesses in a predictable fashion when they have to coordinate with university subjects. For example, foreign-born subjects from countries where lower tips are acceptable would use their knowledge of U.S. norms and rate those same tip amounts as less appropriate when coordinating with other university students. As a result, ratings for foreign-born subjects coordinating with university students should change in a predictable fashion.

Because some of the norms for countries outside of the U.S. are the same as those in the U.S., we can also conduct our analysis on only the sub-sample of foreign subjects whose ex-ante norm is below 12% (this excludes Canada and the Cayman Islands born individuals). We expect these subjects will guess higher levels of appropriateness for lower tip amounts when matched with a same-country born subject than when matched with a university-subject. A paired t-test on this subset of subjects, testing the difference in means of foreign-born nationals' best guess of the norms, uTip and nTip, in the two coordination tasks, found that the mean of uTip–nTip is 3.085 (t(41=4.401, p=0.000) and in the expected direction.

In parallel fashion, we did the same test for punctuality. For foreign-born subjects who come from a country where the ex-ante norm is to arrive 'on-time', the mean difference between their guess about university norms, uPunctual and their guess about their country's punctuality norms, nPunctual, was 0.6730 and is almost statistically significant (t(25)=1.895, p=0.069). For foreign-born subjects whose countries ex-ante norm is to arrive 'late', the mean difference between uPunctual and nPunctual was –2.375 and is not statistically significant (t(19)=–1.538, p=0.140). The difference in the preferred action by matching treatment reference-group is consistent with the identified ex-ante norm, as those matching with 'on-time' countries prefer to arrive earlier than when matching with university-subjects. 'Late' counties, however, display no differences in their preferred arrival time by university-matching or same-county matching treatments.[18] [19]

Overall, these results are consistent with our hypotheses. Personal ratings sometimes differ from those elicited using the coordination tasks, ratings using the coordination task are consistent with ex-ante identified norms and differ depending on the target for the coordination game.

5 Conclusion

In this paper we test for whether the responses from a norm elicitation experiment that uses incentivized coordination games can be appropriately interpreted as an empirical measure

[18] See also supplemental analysis online under http://ekrupka.people.si.umich.edu/

[19] We also conducted a two sample t-test on nPunctual of foreign born subjects where the ex-ante norm is to be "late" and nPunctual of foreign born subjects where the ex-ante norm is to be 'early'. Mean nPunctual of foreign born subjects where the ex-ante norm is to be 'late' is equal to 2.38 where the mean nPunctual of foreign born subjects where the ex-ante norm is to be "early" is equal to 4.52. A two-sided test shows that they are not statistically significant from one another (t(46)=1.472, p=0.148).

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press

of social norms. Furthermore, the responses elicited with the coordination task track ex-ante identified real-world social norms and respondents adjust their coordination game guesses such that they provide different guesses when coordinating with subjects born in the same country and when coordinating with subjects from the same university. Our results provide evidence on the importance of eliciting shared beliefs as a key component of identifying social norms and we offer novel evidence supporting the use of coordination games for eliciting social norms from different reference groups.

References

Akerlof, G. A. & Kranton, R. E. (2005). Identity and the Economics of Organizations. *Journal of Economic Perspectives, 19*(1), 9–32.

Basu, K., & Jörgen, W. (2003). 10 Punctuality: A cultural trait as equilibrium. In *Economics for an Imperfect World: Essays in Honor of Joseph E. Stiglitz*, pp. 163–182. MIT Press.

Bernheim, B. D. (1994). A theory of conformity. *Journal of Political Economy*, *102*(5), 841–877.

Bettenhausen, K. L. & Murnighan, J. K. (1991). The development of an intragroup norm and the effects of interpersonal and structural challenges. *Administrative Science Quarterly*, *36*(1), 20–35.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge, England: Cambridge University Press.

Bicchieri, C. & Chavez, A. (2010). Behaving as expected: Public information and fairness norms. *Journal Behavioral Decision Making*, *23*(2), 161–178.

Bicchieri, C. & Dimant, E. (2019). Nudging with care: The risks and benefits of social information. *Public choice*, 1–22.

Bicchieri, C., Dimant, E., Gächter, S. & Nosenzo, D. (2022). Social proximity and the erosion of norm compliance. *Games and Economic Behavior*, *132*, 59–72.

Bicchieri, C. & Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making, 22*(2), 191–208.

Binmore, K. & Samuelson, L. (2006). The evolution of focal points. *Games and Economic Behavior, 55*(1), 21–42.

Burks, S. V. & Krupka, E. L. (2012). A multimethod approach to identifying norms and normative expectations within a corporate hierarchy: Evidence from the financial services industry. *Management Science*, *58*(1), 203–217.

Camerer, C. F. & Fehr, E. (2004). Measuring social norms and preferences using experimental games: A guide for social scientists. J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, & H. Gintis. (Eds.). *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*, *97*, 55–95.

Catola, M., D'Alessandro, S., Guarnieri, P. & Pizziol, V. (2021). Personal norms in the online public good game. *Economics Letters*, *207*, 110024.

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press

Charness, G., Rigotti, L. & Rustichini, A. (2007). Individual behavior and group membership. *American Economic Review*, *97*(4), 1340–1352.

Cialdini, R. B., Kallgren, C. A. & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In M. P. Zanna (Ed.). *Advances in experimental social psychology, volume 24.* (pp. 201–234). Cambridge, Massachusetts: Academic Press.

Cialdini, R. B., Reno, R. R. & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*(6), 1015–1026.

Crawford, V. P., Gneezy, U. & Rottenstreich, Y. (2008). The power of focal points is limited: Even minute payoff asymmetry may yield large coordination failures. *American Economic Review*, *98*(4), 1443–1458.

Deutsch, M. & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, *51*(3), 629–636.

Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives*, *3*(4), 99–117.

Fehr, E. & Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, *14*(3), 159–181.

Fodor's. (2002). *Fodor's Fyi: How to Tip*. Fodor's Travel Publications.

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., ... & Yamaguchi, S. (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, *332*(6033), 1100–1104.

Gintis, H. (2014). *The bounds of reason*. Princeton University Press.

Gneezy, A., Gneezy, U., Nelson, L. D. & Brown, A. (2010). Shared social responsibility: A field experiment in pay-what-you-want pricing and charitable giving. *Science*, *329*(5989), 325–327.

Goette, L., Huffman, D. & Meier, S. (2006). The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. *American Economic Review*, *96*(2), 212–216.

Harrison, G.W. & Rutström, E. E. (2008). Chapter 81 experimental evidence on the existence of hypothetical bias in value elicitation methods. In C. R. Plott & V. L. Smith (Eds.). *Handbook of experimental economics results, volume 1.* (pp. 752–767). Amsterdam: Elsevier.

Herrbach, O. & Mignonac, K. (2007). Is ethical p–o fit really related to individual outcomes? A study of management-level employees. *Business & Society*, *46*(3), 304–330.

Jones, S. R. (1984). *The economics of conformism.* Blackwell.

Krupka, E. L. & Weber, R. A. (2009). The focusing and informational effects of norms on pro-social behavior. *Journal of Economic Psychology*, *30*(3), 307–320.

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press

Krupka, E. L. & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary?. *Journal of the European Economic Association*, *11*(3), 495–524.

Levine, R. N. (2008). *A geography of time: On tempo, culture, and the pace of life*. New York, New York: Basic Books.

Levine, R. V., West, L. J. & Reis, H. T. (1980). Perceptions of time and punctuality in the United States and Brazil. *Journal of Personality and Social Psychology*, *38*(4), 541.

Lynn, M. (2006). Tipping in restaurants and around the globe: An interdisciplinary review. In M. Altman (Ed.). *Handbook of contemporary behavioral economics: Foundations and development* (pp. 626–643). Armonk, New York: M.E. Sharpe Publishers.

Lynn, M. & Lynn, A. (2004). National values and tipping customs: A replication and extension. *Journal of Hospitality and Tourism Research*, *28*(3), 356–364.

Mehta, J., Starmer, C. & Sugden, R. (1994). The nature of salience: An experimental investigation of pure coordination games. *The American Economic Review*, *84*(3), 658–673.

Miller, D. T. & Prentice, D. A. (1996). The construction of social norms and standards. In E. T. Higgins & A. W. Kruglanski (Eds.). *Social psychology: Handbook of basic principles* (pp. 799–829). New York City, New York: The Guilford Press.

Morrison, T., Conaway, W. A., Borden, G. A., & Koehler, H. (1994). Kiss, bow, or shake hands: How to do business in sixty countries (p. 456). Holbrook, Mass.: Adams Media Corporation.

Opp, K. D. (2001). How do norms emerge? An outline of a theory. *Mind & Society*, *2*(1), 101–128.

Posner, R. A. & Rasmusen, E. B. (1999). Creating and enforcing norms, with special reference to sanctions. *International Review of law and economics*, *19*(3), 369–382.

Reuben, E. & Riedl, A. (2013). Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior*, *77*(1), 122–137.

Roth, A. E. & Kagel, J.H. (2016). *The handbook of experimental economics, volume 2*. Princeton: Princeton University Press.

Schelling, T. C. (1980). *The strategy of conflict: with a new preface by the author*. Cambridge, Massachusetts: Harvard University Press.

Schram, A. & Charness, G. (2015). Inducing social norms in laboratory allocation choices. *Management Science*, *61*(7), 1531–1546.

Schwartz, S. H. (1968). Awareness of consequences and the influence of moral norms on interpersonal behavior. *Sociometry*, *31*(4), 355–369.

Schwartz, S. H. (1973). Normative explanations of helping behavior: A critique, proposal, and empirical test. *Journal of Experimental Social Psychology*, *9*(4), 349–364.

Schwartz, S. H. (1977). Normative influences on altruism. In *Advances in Experimental Social Psychology* (Vol. 10, pp. 221–279). Academic Press.

Star, N. (1988). The international guide to tipping. Berkley Publishing Group.

Sugden, R. (1995). A theory of focal points. *The Economic Journal*, *105*(430), 533–550.

Sugden, R. (2000). The motivating power of expectations. In J. Nida-Rümelin & W. Spohn. (Eds.). *Rationality, rules, and structure*. (pp. 103–129). Dordrecht: Springer.

White, L. T., Valk, R. & Dialmy, A. (2011). What is the meaning of "on time"? The sociocultural nature of punctuality. *Journal of Cross-Cultural Psychology*, *42*(3), 482–493.

World Travelers of America. (n.d.). *Worldwide tipping guide*. World Travelers of America: Worldwide Tipping Guide. Retrieved February 21, 2022, from https://worldtravelers.org/travel-tips-tipping-guide.asp.

Young, P. H. (2007). Social norms. (Economics Series Working Papers 307). University of Oxford, Department of Economics.

https://doi.org/10.1017/S1930297500009104 Published online by Cambridge University Press