Measuring social norm variation across contexts: Replication and comparison to alternative methods

David Huffman, Garrett Kohno, Pauline Madiès, Spencer Vogrinec,
Stephanie W. Wang, Dhwani Yagnaraman*

May 2025

Abstract

Studying social norms and how they vary in different contexts requires reliable measurements. We revisit the seminal Krupka and Weber (KW 2013) norm elicitation method and assess the importance of two dimensions of eliciting norms — first, whether to use the KW coordination game method or an alternative, two-stage method that directly elicits first-order or second-order beliefs about social appropriateness, and second, whether to use financial incentives. We replicate KW's main finding of a qualitative difference in norms between the dictator game and a re-framed version that involves potentially taking money: KW and all other methods show that taking money is less socially appropriate than giving money, holding outcomes fixed and regardless of the presence of monetary incentives. However, we find that the difference in elicited social appropriateness between the two versions of the dictator game varies across methods, with elicited first-order beliefs exhibiting the largest gap in social appropriateness and KW exhibiting the smallest gap. One possible explanation is that strategic uncertainty and complexity in the KW method may attenuate sensitivity of the measure to differences in norms across contexts. A comprehension check reveals that about half of the KW participants initially misunderstood the task, and a prediction exercise reveals that first-order beliefs yield the best predictive power over actual behavior in simple dictator games. One implication is that first-order beliefs could be a simple alternative measure for capturing norm differences across contexts, with good predictive power. A caveat, however, is that first-order beliefs might be more subject to social desirability bias in settings with controversial or pluralistic norms.

JEL Classifications: C91, D64, D91
Keywords: Social Norms; Norm Elicitation; Incentives; Higher-Order Beliefs

1 Introduction

Social norms are often defined as the commonly known standard of behavior that is based on widely shared views of how individual group members ought to behave in a given situation (Elster, 1989; Bicchieri, 2006). These incorporate both expectations of how others ought to behave (injunctive norms) and how others actually behave (descriptive norms) (Cialdini et al., 1990; Bicchieri, 2006). Prior studies suggest that social norms play an important role in many areas such as prosocial behavior (Krupka & Weber, 2013; Bicchieri et al., 2022), honesty (Abeler et al., 2019; Bicchieri et al., 2023), discrimination (Barr et al., 2018), and female labor force participation (Bursztyn et al., 2020), to name a few.

Studying social norms requires reliable measurements, and Krupka and Weber (2013) made a seminal contribution in this regard. Their approach to measuring injunctive norms (henceforth the KW method) asks participants to rate the social appropriateness of different actions in experimental games and incentivizes them to guess the modal response of other participants. The core result of their study is that participants indicated different levels of social appropriateness for actions across two games: the standard dictator game and a re-framed version called the "bully" game, in which different initial endowments allowed dictators to "take" money from recipients. Beyond identifying normative differences between the two environments, KW demonstrate how these incentivized norms can both explain and predict behavioral variation across the games.

One goal of our paper is to provide a replication test of the main result of Krupka and Weber (2013). We implement the same two versions of the dictator game, using the same parameters and instructions, with

*We thank Ben Greiner for the inspiration. We thank Luca Braghieri, Erin Krupka, Roberto Weber, and conference and seminar participants at the External Validity, Generalizability and Replicability of Economics Experiment Workshop of the 2024 Barcelona Summer Forum, NHH-Rady Spring School in Behavioral Economics, and the University of Pittsburgh Experimental Economics Brown Bag for their helpful comments. This study was pre-registered as AEARCTR-0012061.
Huffman: Department of Economics, University of Pittsburgh, Pittsburgh, PA, huffmand@pitt.edu; Kohno: Department of Economics, University of Pittsburgh, Pittsburgh, PA gkh8@pitt.edu; Madiès: Department of Economics, Sciences Po, Paris, France, pauline.madies@sciencespo.fr; Vogrinec: Katz Graduate School of Business, University of Pittsburgh, PA, spencer.vogrinec@pitt.edu; Wang: Department of Economics, University of Pittsburgh, PA, swwang@pitt.edu; Yagnaraman: Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA, dyagnara@andrew.cmu.edu

a notable difference being that we used a representative online sample instead of college students.[1] To our knowledge, we provide the first attempt to replicate this aspect of the original study.

Another goal of our paper is to evaluate the KW method relative to other methods in detecting differences in norms across games. One reason for testing alternative methods is the conceptual ambiguity surrounding the beliefs measured by the KW method. In particular, injunctive norms are usually conceptualized as second-order beliefs, or beliefs that an individual holds about what others believe is appropriate or inappropriate in a given context (Görges and Nosenzo, 2020). Using a coordination game, the KW method incentivizes participants to guess what others guess, but since everyone faces the same incentive, higher-order beliefs potentially come into play (participants need to guess what other participants guess about what other participants guess, etc.). Incentivizing the formation of higher-order beliefs may introduce strategic uncertainty into the task. Furthermore, in the original implementation of the method, the instructions ask for personal beliefs about the social appropriateness of an action, but the incentive is to indicate the most common response among all participants in the session or study. This dissonance may create confusion due to poor understanding of the task (König-Kersting 2024). Either of these aspects of the KW method could increase uncertainty in subjects' optimal response and attenuate the measurement of how norms vary across different environments.

Alternative, two-stage versions of norm measures have been proposed, most notably Bicchieri and Xiao (2009), where respondents initially state their opinion about what is the most appropriate action in a given context, and a second stage elicits guesses of the most common personal normative belief. Such methods may be more sensitive to detect norm differences across games by eliminating strategic uncertainty or general confusion. We implement a method consisting of two stages, where one group of participants, the "evaluators," initially provide personal evaluations of the social appropriateness of a set of given actions (first-order beliefs)[2], and in the second stage other participants guess the most common first-stage beliefs for the same set of actions (second-order beliefs). One potential drawback of two-stage methods is the lack of incentives in the evaluator beliefs, which may introduce social desirability bias (Aycinena et al. 2024).

Another important methodological question we investigate is how crucial incentives are for delivering reliable measures of norm differences across games. We provide a first comparison of the KW method versus two-stage methods, in the presence and absence of incentives, in terms of their ability to detect differences in norms across and predict behavior across the standard and bully versions of the dictator game.

The first stage of our analysis shows that the core result of Krupka and Weber (2013) is replicated. Across dictator and bully games, there is a significant difference in responses using the KW method. Participants indicated higher levels of social appropriateness for actions that involve "not giving" in the dictator game, compared to actions involving "taking" in the bully game. This difference and also the magnitudes of the appropriateness ratings in each game are quite similar to those found in the original study.

The second stage of our analysis compares how participants respond to different measures of social norms in dictator and bully games. We find that all measures deliver the same qualitative result on the difference in norms across games, regardless of monetary incentives. However, we find that the magnitude of differences in social appropriateness between the two choice environments varies across methods; the participants who stated their personal social appropriateness ratings directly in the first stage (the Evaluators) exhibit the largest gap in social appropriateness, and the KW methods exhibit the smallest gap. One potential explanation is that greater uncertainty about how others will respond leads to attenuation in sensitivity to changes in the contextual features of the game.[3] Another source of attenuation in the KW method could be confusion due to the dissonance between the task instructions and the incentive. A comprehension check prior to the dictator game suggests that participants have difficulty understanding their task under the KW method.

The last stage of our analysis implements a prediction exercise where we predict the actual behavior observed in the original KW data using our five methods. We find that the Evaluator method best predicts actual choices made in the dictator games. However, results of our prediction exercise should be taken with caution. While the Evaluator method may be suitable to measure norms and use them to predict behavior in a simple dictator game setting, there is reason to be cautious about generalizing this result to other settings. In particular, the method might not be suitable in more complex settings with pluralistic or controversial

[1] We also included a comprehension question between the instructions and the experiment choices, which tests awareness of the rules for incentivization.

[2] We use the term first-order beliefs of social appropriateness to refer to an individual's beliefs about what society views as appropriate behavior. It is conceptually distinct from personal normative beliefs (Bicchieri, 2006), personal norms (Bašić & Verrina, 2024), and personal beliefs (Barigozzi & Montinari, 2023), which emphasize an individual's own sense of appropriateness, often independent of others' views. In contrast, first-order beliefs in our setting cannot be separated from the perceptions of others. Our definition also differs from the notion of first-order beliefs in game theory, where the term refers to beliefs about other players' actions (see, e.g., Fudenberg & Tirole, 1991).

[3] In environments where multiple norms exist, this may be compounded by inherent normative uncertainty. (Fromell et al., 2021; Dimant, 2023; Dimant et al. 2024, Aycinena et al., 2024; Panizza et al., 2024; Kimbrough et al., 2024)

norms, because it cannot be incentivized, and thus might be prone to social desirability bias.

Our paper contributes to a previous literature that evaluates the KW method and compares it to alternative methods of measuring social norms. d'Adda et al. (2016) show that the KW method is robust to order effects in a bribery game, and Veselý (2015) finds similar appropriateness ratings in an ultimatum game in the absence of incentives when using the KW method. Fallucchi and Nosenzo (2022) raise the possibility that salient focal points could skew answers in the KW method due to the presence of multiple equilibria and find that the KW method is robust to the inclusion of visual labels, except when there is no clear norm. Aycinena et al. (2024) compare the predictive power of the KW method, the Bicchieri and Xiao method, and a novel binarized scoring method and find that the KW method yields the norms that are most predictive of observed behavior for variations of the dictator game that don't include the bully game. König-Kersting (2024) tests whether modifications to instructions designed to change the salience of aspects of the coordination game and monetary incentives affect responses under the KW method. The study also compares the KW method with a two-stage method and finds that none of these variations yields significant differences in elicited norms in the standard dictator game. Our study differs from these studies in that we offer a replication of the main finding in Krupka & Weber (2013) and evaluate how alternative elicitation methods capture norm differences between the standard dictator game and the bully variant. We find small differences across measurement approaches for norms in the standard dictator game but find a substantial difference across methods when comparing the gap in elicited norms for the standard dictator game to norms for the bully game, consistent with increasing attenuation as methods become progressively more complex.

This paper is organized as follows. Section 2 describes our online experiment that tests five elicitation methods. Section 3 presents results. Section 4 concludes.

2 Experiment

2.1 Design

Table 1 summarizes the treatment conditions, which vary the type of beliefs elicited and use of monetary incentives.[4] The treatment conditions are administered between-subjects. Participants in the *Evaluators* condition are instructed to submit a personal evaluation about the social appropriateness (first-order beliefs) of taking certain actions in the dictator game. Participants in the *Second-Stage* (*Non-Incentivized Second-Stage*) condition are instructed to guess the actual (non-incentivized) modal belief of the *Evaluators* by forming second-order beliefs. Participants in the *Krupka-Weber* condition are given the same instructions as the *Evaluators*, except they are incentivized to guess the modal belief of others in the study, thereby forming higher-order beliefs. Finally, participants in the *Non-Incentivized Krupka-Weber* condition are instructed to guess the modal belief of others in the study, again forming higher-order beliefs.[5]

Table 1: Experimental Conditions

Treatment	**Elicited Beliefs**	**Monetary Incentive**	**N**
Evaluators	First-order	No	151
Second-Stage	Second-order	Yes	153
Non-Incentivized Second-Stage	Second-order	No	148
Krupka-Weber	Higher-order	Yes	152
Non-Incentivized Krupka-Weber	Higher-order	No	147

Within each elicitation method, participants are randomly assigned to receive the "standard" or the "bully" choice environment of the dictator game from Krupka & Weber (2013). In the standard game, participants are presented with a scenario in which a dictator is endowed with $10 and can choose to give any amount of this money, in one-dollar increments to the recipient, who initially receives $0. An example of an action might be for Individual A (dictator) to "Give $4 to Individual B. Individual A gets $6, Individual

[4] The instructions for each condition's task can be found in Table A.1 in Appendix A.1.

[5] Our Krukpa-Weber condition is a close replication of Krupka and Weber (2013) (Brandt et al., 2014; Chen et al., 2021). There are two primary differences between our Krupka-Weber condition and the original KW method. The first difference is that we administer the experiment with an online sample rather than a student sample in-person. The second difference is that we include a comprehension check before participants indicate their social appropriateness ratings. One minor difference is that participants were paid a $3 participation fee and a $2 additional payment if they correctly guessed the modal response in incentivized conditions whereas KW paid a $7 participation fee and either a $5 or $10 additional bonus for guessing the modal response. We view this last difference as minor due to acceptable participant payments online being lower than acceptable payments in-person.

B gets \$4," which is then represented as action (\$6, \$4). In the bully game, participants are presented with a scenario in which both the dictator and the recipient initially receive \$5. In this choice environment, the dictator has the opportunity to not only *give*, but also to *take* money from the recipient in one-dollar increments.

The two alternative choice environments are such that the set of potential payoffs is the same, while the contextual features of the action sets differ. To illustrate this, the corresponding action in the bully game that leads to the same outcome in the previous example of the standard game is for Individual A (dictator) to "Take \$1 from Individual B. Individual A gets \$6, Individual B gets \$4," which is then represented as action (\$6, \$4). Although both actions yield the same payoffs, they may be governed by different social norms, with actions involving taking generally considered less socially appropriate than actions involving giving (Krupka & Weber 2013).

2.2 Procedure

The experiment was conducted on the Prolific survey platform using a representative US sample balanced on age, sex, and political affiliation. The sample consisted of 751 participants. Data collection occurred in January 2024. The experiment was coded in oTree (Chen et al. 2016).

The experiment begins with a set of instructions that vary by elicitation method (see Table A.1). To ensure that participants understand the instructions, an example situation is provided. After participants respond to the example situation, they receive an explanation of how they could have responded based on their elicitation method.

The example is followed by a comprehension question, which simply asked participants how they should indicate their responses.[6] The possible answer options displayed for this question differ based on the assigned elicitation method. All participants see two options: (i) reporting personal beliefs about social appropriateness and (ii) an obviously incorrect option to ensure that they are paying attention. Those assigned to the *Second-Stage* or *Krupka-Weber* elicitation methods (including the non-incentivized versions) also see a third option about reporting the most frequently given rating by evaluators, or all participants in the study, respectively. This differentiation ensures that participants in the KW and Second-Stage conditions do not become aware of the other elicitation method and do not mistakenly attempt the evaluator task. Participants answer the comprehension question as many times as needed to get the correct answer and proceed.

Participants are then presented with the main norm elicitation task for either the standard or bully variant of the dictator game. In one of the choice environments, participants rate the social appropriateness of each possible action on a four-point scale ranging from "Very socially inappropriate" to "Very socially appropriate",[7] in line with their assigned elicitation method. After participants indicated social appropriateness ratings for all eleven possible dictator actions, the experimenter randomly selected one possible action in each variant of the game. Participants in the *Second-Stage* and *Krupka-Weber* conditions received a bonus payment of \$2, in addition to a \$3 participation payment that all participants received, if they had selected the modal appropriateness rating in the Evaluators condition or the Krupka-Weber condition, respectively, for the selected action.[8]

3 Results

Our data contains 751 participants. The average age of the sample is 45.79 and 49% of all participants are female. Table 2 shows summary statistics for all elicitation methods.

3.1 Identifying taking vs. giving norms

First, we investigate whether the two dictator game variants are governed by a different set of norms. We expect actions that leave the dictator with more money than the recipient, namely outcomes (\$10, \$0) to (\$6, \$4), to be associated with lower social appropriateness ratings in the bully game than in the standard game. These actions involve *taking* money from the recipient in the bully game, which is considered less socially appropriate than actions in the standard game that involve *giving* money to the recipient.

Table 3 compares social appropriateness ratings elicited by the different methods in the two choice environments. The first column reports differences in social appropriateness ratings across games and methods for outcomes (\$10, \$0) to (\$6, \$4). We replicate Krupka & Weber's (2013) findings of a difference in norms

[6] For incentivized conditions the comprehension question specified "How should you indicate your responses to earn a bonus of \$2?".

[7] The scale comprises four possible ratings: "Very socially inappropriate"(coded as -1), "Somewhat socially inappropriate" (coded as -0.3), "Somewhat socially appropriate" (coded as 0.3) and "Very socially appropriate" (coded as 1).

[8] Across all treatments, the average earnings were \$3.73, while the average earnings per hour were \$16.54.

Table 2: Summary Statistics

	Evaluators	NI Second-Stage	Second-Stage	NI Krupka-Weber	Krupka-Weber
	(1)	(2)	(3)	(4)	(5)
Bully treatment	0.53	0.47	0.52	0.46	0.41
Avg. incorrect on comprehension	0.07	0.24	0.16	0.55	0.39
Age	44.38	45.79	47.65	45.99	45.13
Female	0.54	0.49	0.52	0.48	0.45
Race					
Asian	0.11	0.06	0.08	0.09	0.07
Black	0.13	0.15	0.13	0.12	0.07
Caucasian	0.70	0.71	0.71	0.73	0.74
Hispanic	0.05	0.05	0.05	0.03	0.08
Other	0.01	0.03	0.03	0.03	0.05
Moved to USA					
Born in the USA	0.91	0.95	0.92	0.93	0.93
Before 5	0.02	0.01	0.01	0.01	0.02
Age 5-10	0.01	0.01	0.01	0.00	0.00
Age 11-13	0.01	0.01	0.01	0.01	0.00
Age 14-18	0.02	0.01	0.00	0.01	0.02
After 18	0.04	0.02	0.05	0.04	0.03
Political Affiliation					
Democrat	0.32	0.39	0.28	0.30	0.30
Republican	0.29	0.28	0.23	0.27	0.23
Independent	0.38	0.33	0.48	0.44	0.47
Observations	151	148	153	147	152

Notes: For each variable, the columns display the mean. Column 1 shows these statistics for the evaluators method, column 2 for the non-incentivized second-stage method, column 3 for the second-stage method, column 4 for the non-incentivized Krupka-Weber method, and column 5 for the Krupka-Weber method.

between "standard" and "bully" dictator games. Robust to all elicitation methods, taking money is considered less socially appropriate than giving money to the recipient when the two choices result in the same monetary outcome. Indeed, social appropriateness of the taking environment is significantly less than that of the giving environment in outcomes ($10, $0) to ($6, $4) and not statistically different for outcomes ($5, $5) to ($0, $10).

3.2 Incentives

Table 4 reports the effect of monetary incentives on social appropriateness ratings. Across both games and all outcomes, the coefficient on incentives is not different from 0, suggesting that monetary incentives do not yield any significant differences in elicited norms between methods. Additionally, there are no systematic differences between the variances of ratings between the incentivized and non-incentivized methods.[9]

3.3 Gap in Social Appropriateness

Our results suggest that all methods identify a difference in social norms for outcomes ($10, $0) to ($6, $4), where taking is uniformly considered less socially appropriate than giving. However, a difference in the magnitude of this gap in social appropriateness emerges between elicitation methods. Figure 1 shows the social appropriateness ratings for the two games by elicitation method.[10] The difference in norm ratings across games is the largest when we elicit first-order beliefs, namely the judgments about social appropriateness of the Evaluators, and the smallest when we elicit higher-order beliefs using the KW method. The difference in the gaps is statistically significant for outcomes ($10, $0) to ($6, $4), as reported in Table 3. A post-hoc power analyses with observed means show that estimated power ranges from 0.83 to 1.00 for these comparisons.

Attenuation of the taking-giving gap under the KW method could be due to judging each action on the resulting monetary payoffs, rather than on other contextual features of the games. This might arise from strategic uncertainty or confusion. There is inherent strategic uncertainty in the KW methods as participants guess each others' guesses about each others' guesses, rather than simply guessing what others

[9]Table A2 in the Appendix A.2 presents the results of F-tests comparing the variance in ratings of incentivized and non-incentivized participants for both games and all outcomes.

[10]The incentivized and non-incentivized versions of the KW method and Two-Stage method are pooled together for this figure since the presence of the incentive does not significantly affect responses. A version of the figure showing each condition separately is available in Figure A1 in the Appendix A.3.

Table 3: Difference in Giving and Taking Norms

	Allocation Outcomes	
	($10,$0) - ($6,$4)	($5,$5) - ($0,$10)
Bully	-0.221***	0.066
	(0.059)	(0.087)
Bully × Evaluators	-0.205**	-0.149
	(0.096)	(0.126)
Bully × NI Second-Stage	-0.049	0.091
	(0.085)	(0.122)
Bully × Second-Stage	-0.140	0.000
	(0.094)	(0.123)
Bully × NI Krupka-Weber	0.055	-0.118
	(0.088)	(0.126)
Evaluators	0.168**	0.110
	(0.074)	(0.091)
NI Second-Stage	0.014	-0.096
	(0.059)	(0.086)
Second-Stage	0.141*	0.055
	(0.072)	(0.088)
NI Krupka-Weber	0.020	0.011
	(0.063)	(0.090)
Constant	-0.420***	0.297***
	(0.041)	(0.064)
Observations	3,755	4,506
R^2	0.07	0.01

Notes: The dependent variable for both columns are on a 4-point Likert scale where -1 is "Very socially inappropriate" and 1 is "Very socially appropriate." The reference group for the allocation game is the "Standard" variant. The reference group for the elicitation method is *Krupka-Weber*. OLS estimates are presented with robust standard errors in parentheses. Standard errors are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

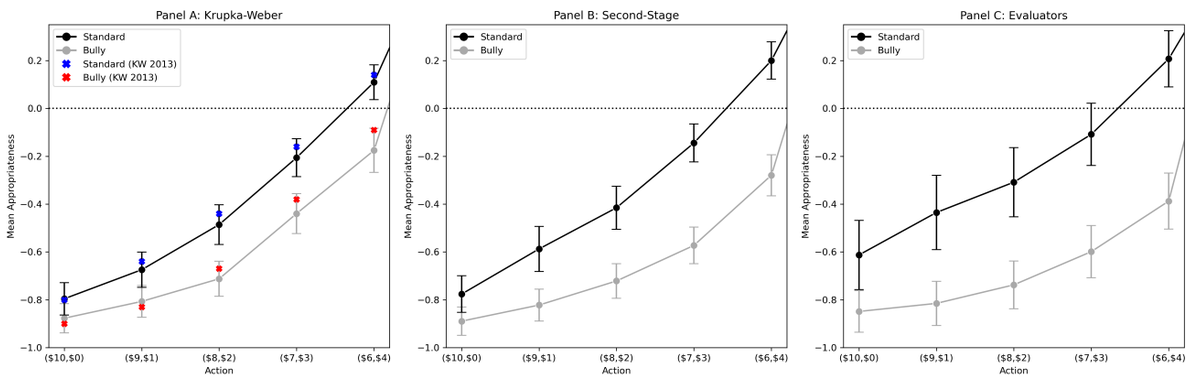
Table 4: Presence of incentives on dictator game norms

	Standard		Bully	
	($10,$0) - ($6,$4)	($5,$5) - ($0,$10)	($10,$0) - ($6,$4)	($5,$5) - ($0,$10)
Second-Stage	-0.006	-0.107	-0.110*	0.101
	(0.064)	(0.085)	(0.062)	(0.092)
Incentivized	-0.020	-0.011	-0.074	0.107
	(0.063)	(0.090)	(0.061)	(0.088)
Second-Stage × Incentivized	0.146	0.162	0.110	-0.047
	(0.096)	(0.122)	(0.086)	(0.126)
Constant	-0.400***	0.308***	-0.566***	0.256***
	(0.047)	(0.063)	(0.044)	(0.066)
Observations	1,595	1,914	1,405	1,686
R^2	0.01	0.01	0.01	0.01

Notes: The dependent variable for all columns are on a 4-point Likert scale where -1 is "Very socially inappropriate" and 1 is "Very socially appropriate." The reference group for the elicitation method is *Non-Incentivized Krupka-Weber*. OLS estimates are presented with robust standard errors in parentheses. Standard errors are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

think is socially appropriate. The uncertainty present in others' responses could lead to attenuation through the deliberate use of monetary payoffs as a coordination device. Furthermore, the dissonance between the task instructions and incentives under the KW method may lead to confusion about how to perform the task. Participants are asked to state first-order beliefs but are incentivized to match the modal response in their responses. Therefore, they might be confused about which level of beliefs to report. Both aspects of the KW method contribute to uncertainty over one's own optimal decision, which has been linked to systematic attenuation of economic decisions and compressing toward intermediate cognitive defaults (Enke & Graeber 2023).

Figure 1: Giving and Taking Norms by Elicitation Method



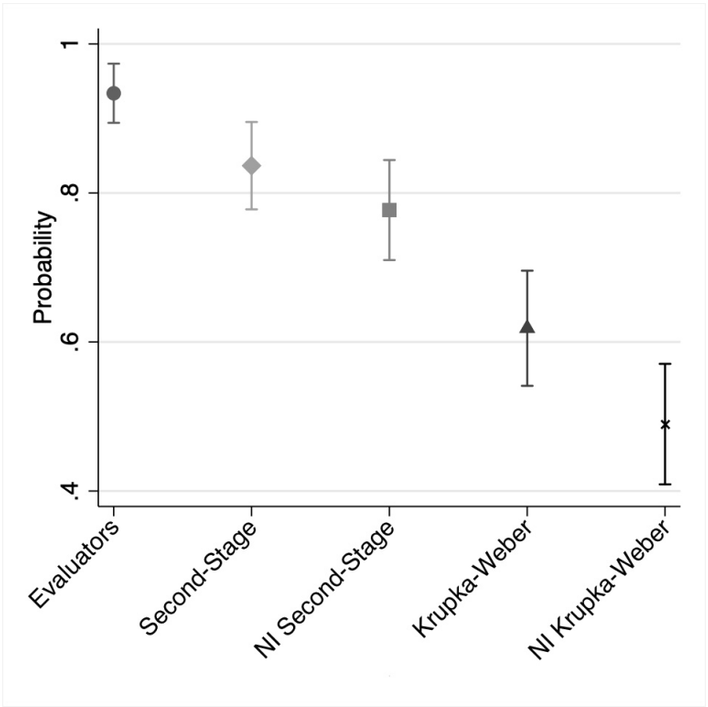
Notes: This figure shows the mean appropriateness ratings of each action in the standard and bully environments of the dictator game for outcomes ($10,$0) to ($6,$4). Panel A uses data from the Krupka-Weber elicitation method, pooling data from the incentivized and non-incentivized conditions. It also shows the corresponding ratings from the original Krupka & Weber (2013) study. Panel B uses data from the Second-Stage elicitation method, pooling data from the incentivized and non-incentivized conditions. Panel C uses data from the Evaluator condition. Error bars reflect 95 percent confidence intervals.

3.4 Comprehension

To further investigate the role of confusion, we examine the comprehension check, which consists of one multiple choice question asking participants how they should respond in the subsequent norm elicitation task. Figure 2 reports the estimated probability of passing the comprehension check on the first attempt by treatment, using a probit regression. Almost 50% of participants fail the comprehension check on their initial attempt in the KW methods, compared to less than 10% for the evaluators. As anticipated, the Second-Stage methods are simpler for participants to understand than the KW methods, but more complex than being asked for first-order beliefs about social appropriateness (Evaluators). Our comprehension check requires participants to demonstrate their understanding of the task before starting the task, unlike Krupka-Weber (2013) where there was no test of task comprehension. Despite this, Figure 1 shows few differences between the average norm ratings in our KW methods and the original paper.[11] Poor understanding of the task in the KW methods is consistent with König-Kersting (2024), who finds that including a mandatory comprehension check in the pre-task instructions improves post-task recall of the task and incentives, but also finds no significant differences in elicited norms. Although there is substantial confusion with the task at hand in KW methods, the severity of the problem is unclear as attempts to reduce dissonance and confusion have not produced meaningful differences in norms.

[11] Comparing our *Krupka-Weber* incentivized method to the original data, only 3/22 norms are significantly different at the 0.05 level using rank-sum tests.

Figure 2: Initial task comprehension by method



Notes: This figure shows Probit estimates for the probability of passing the comprehension check question on the first atttempt, by method. Error bars reflect 95 percent confidence intervals.

3.5 Predictive Power

Beyond identifying differences in social appropriateness ratings, we evaluate the methods on their ability to predict differences in actual behavior across games. Using data on real behavior from Krupka & Weber (2013), we estimate the following norm-dependent utility model across methods:

$$U(a_j) = \beta\pi(a_j) + \gamma N(a_j) \tag{1}$$

where $\pi(a_j)$ is the monetary payoff associated with each of the $j = 1, ..., 11$ actions a_j in a Dictator Game and $N(a_j)$ is the social appropriateness associated with each of these actions. While $\pi(a_j)$ remains constant across treatments, $N(a_j)$ reflects variation in the mean social appropriateness of action a_j across elicitation methods. Our parameters of interest, β and γ, capture the extent to which participants weigh monetary payoffs (β) and social norms (γ) when choosing an action. For each elicitation method, we combine behavioral data from both versions of the dictator game used in Krupka & Weber (2013)[12] with the average social appropriateness rating for each action under each method. We then estimate β and γ using a conditional logit model, where the binary outcome indicates whether a given action was chosen.

Using estimates of β and γ for each method, we compute the probabilities of choosing each action and compare how well the predictions fit actual choices. Figure 3 shows observed choices and predicted choices under each method in the dictator games. All methods correctly predict that more subjects will choose the equal split ($5, $5) in the bully game than in the standard game, and that conditional on not selecting the equal split, more subjects will select the payoff maximizing option ($10, $0) in the bully game than the standard game. We test which method yields norms that best fit the observed data using the Akaike information criterion (AIC). As reported in the table below, the evaluator method yields the lowest AIC score, suggesting that this method provides the best data fit. For the Second-Stage and Krupka-Weber methods, it is inconclusive whether one fits the data significantly better than the other. Although the first-order beliefs elicited in the Evaluator condition identify the largest normative difference between taking and giving and outperform the other methods in predicting observed behavior, we recommend exercising caution when using them to measure social norms.

Respondents are asked to evaluate what is "socially" appropriate, defined for them as behavior that most people agree is the "correct" or "ethical" thing to do. Despite explicitly requesting that respondents incorporate others' views into their first-order beliefs, it differs from the usual measures of injunctive norms which directly incentivize the formation of beliefs about others' beliefs. First-order beliefs might therefore reflect personal opinions of appropriateness to some extent. This could not only undermine them as a

[12]The data come from Experiment 2 in Krupka & Weber (2013), which includes 106 students from Carnegie Mellon University: 52 in the standard treatment and 54 in the bully treatment.

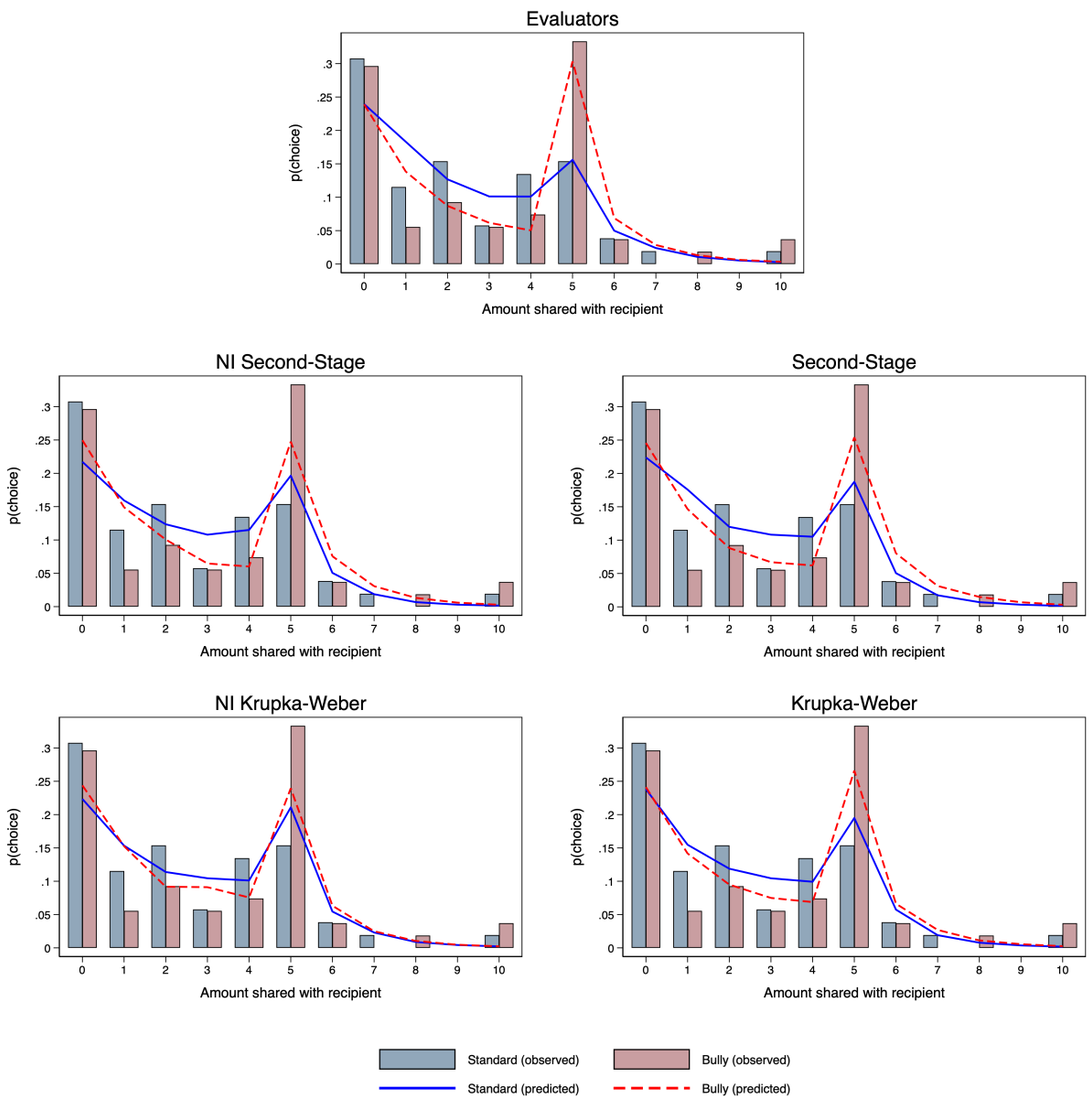
measure of normative expectations, but also render them more susceptible to social desirability bias. This concern is amplified in environments where multiple or contested norms coexist, as personal views are more likely to diverge from the modal or average belief in the population. Thus, while effective in our setting where there is likely considerable overlap between the two, this measure may not generalize well to more complex environments with controversial or pluralistic norms.

Table 5: AIC and BIC scores for each treatment

Method	**AIC**	**BIC**
Evaluators	417.77	423.10
Second-Stage	421.47	426.80
Non-Incentivized Second-Stage	421.29	426.62
Krupka-Weber (our data)	419.63	424.95
Non Incentivized Krupka-Weber	422.51	427.83
Krupka-Weber (2013 data)	420.91	426.23

Notes: This table shows how well the norms from each elicitation method fit the observed data using AIC scores.

Figure 3: Predicted and observed choices in dictator games



Notes: This figure shows observed choices (bars) and predicted choices (lines) in each variant of the dictator game. Each panel corresponds to an experimental condition and shows the predicted choices using the norms elicited under that method.

4 Conclusion

Our study assesses two dimensions of eliciting social norms: the use of financial incentives, and the choice between the coordination game approach of Krupka & Weber (2013) and a two-stage method that directly

measures first- and second-order beliefs separately. We implement five methods that differ in the type of beliefs they elicit and in the use of monetary incentives. First, we assess these methods for their ability to identify a difference in social norms where we expect it to exist. Using a standard dictator game and a variant with differing initial endowments ("Bully Game"), we replicate KW's finding of a qualitative difference in norms between these games. All methods show that taking money is less socially appropriate than giving money, holding the outcomes fixed, regardless of the presence of monetary incentives. We find that the difference in social appropriateness between the standard and bully dictator games varies across methods, with the Evaluators (first-order beliefs) exhibiting the largest gap in social appropriateness, and the KW methods exhibiting the smallest gap. A comprehension check before norm elicitation reveals that nearly half of KW participants initially misunderstood the task. Results from a prediction exercise suggest that first-order beliefs from the Evaluators best predict actual behavior in the original KW data. Our results suggest that complex norm elicitation methods may be less sensitive to the norm-relevant context than measuring first- or second-order beliefs directly due to attenuation arising largely from strategic uncertainty rather than confusion. Due to this, the KW method provides a conservative test for uncovering differences in social norms across different environments. In addition, while first-order beliefs outperform the other methods in identifying norms and predicting behavior in our simple dictator games, it may not be a suitable measure of norms in other, more complex settings with pluralistic or controversial norms.

Support

The authors are grateful for financial support from the Behavioral Economics Design Initiative at the University of Pittsburgh. Wang completed this work as a Fellow at the Center for Advanced Study in the Behavioral Sciences.

References

[1] Abeler, J., Nosenzo, D., & Raymond, C. (2019). Preferences for truth-telling. *Econometrica, 87*(4), 1115-1153.

[2] Aycinena, D., Bogliacino, F., & Kimbrough, E. O. (2024). Measuring norms: a comparison of the predictive and descriptive power of three methods. *Available at SSRN 4663919*.

[3] Barigozzi, F., & Montinari, N. (2023). Social norms: Personal beliefs versus normative expectations. *Available at SSRN 4989968*.

[4] Barr, A., Lane, T., & Nosenzo, D. (2018). On the social inappropriateness of discrimination. *Journal of Public Economics, 164,* 153-164

[5] Bašić, Z., & Verrina, E. (2024). Personal norms—and not only social norms—shape economic behavior. Journal of Public Economics, 239, 105255.

[6] Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

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

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

[9] Bicchieri, C., Dimant, E., Gelfand, M., & Sonderegger, S. (2023). Social norms and behavior change: The interdisciplinary research frontier. *Journal of Economic Behavior & Organization, 205*, A4-A7.

[10] Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... & Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication?. *Journal of Experimental Social Psychology*, *50*, 217-224.

[11] Bursztyn, L., González, A. L., & Yanagizawa-Drott, D. (2020). Misperceived social norms: Women working outside the home in Saudi Arabia. *American economic review, 110*(10), 2997-3029.

[12] Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance, 9*, 88-97.

[13] Chen, R., Chen, Y., & Riyanto, Y. E. (2021). Best practices in replication: a case study of common information in coordination games. *Experimental Economics*, *24*, 2-30.

[14] 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.

[15] d'Adda, G., Drouvelis, M., & Nosenzo, D. (2016). Norm elicitation in within-subject designs: Testing for order effects. *Journal of Behavioral and Experimental Economics*, *62*, 1-7.

[16] Dimant, E. (2023). Beyond average: a method for measuring the tightness, looseness, and polarization of social norms. *Economics Letters*, *233*, 111417.

[17] Dimant, E., Gelfand, M., Hochleitner, A., & Sonderegger, S. (2024). Strategic behavior with tight, loose, and polarized norms. *Management Science*.

[18] Elster, J. (1989). Social norms and economic theory. Journal of Economic Perspectives, 3(4), 99-117.

[19] Enke, B., Graeber, T. (2023). Cognitive Uncertainty. *The Quarterly Journal of Economics, 138*(4), 2021–2067.

[20] Fromell, H., Nosenzo, D., Owens, T., & Tufano, F. (2021). One size does not fit all: Plurality of social norms and saving behavior in Kenya. *Journal of Economic Behavior & Organization*, *192*, 73-91.

[21] Fallucchi, F., & Nosenzo, D. (2022). The coordinating power of social norms. *Experimental Economics, 25*(1), 1-25.

[22] Fudenberg, D., & Tirole, J. (1991). *Game theory*. MIT press.

[23] Görges, L., & Nosenzo, D., (2020). Measuring social norms in economics: Why it is important and how it is done. *Analyse & Kritik, 42*(2), 285-312.

[24] Kimbrough, E. O., Krupka, E. L., Kumar, R., Murray, J. M., Ramalingam, A., Sánchez-Franco, S., ... & Hunter, R. F. (2024). On the stability of norms and norm-following propensity: a cross-cultural panel study with adolescents. *Experimental Economics*, *27*(2), 351-378.

[25] König-Kersting, C. (2024). On the robustness of social norm elicitation. *Journal of the Economic Sciences Association.*

[26] 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.

[27] Panizza, F., Dimant, E., Kimbrough, E. O., & Vostroknutov, A. (2024). Measuring norm pluralism and perceived polarization in US politics. *PNAS nexus*, *3*(10), pgae413.

[28] Veselý, S. (2015). Elicitation of normative and fairness judgments: Do incentives matter?. *Judgment and Decision making*, 10(2), 191-197.

A Appendix

A.1 Instructions for each elicitation method

Table A1: Instructions for main task in each condition

Treatments	**Instructions**
Evaluators	*For each of the choices, please indicate whether you believe choosing that option is very socially inappropriate, somewhat socially inappropriate, somewhat socially appropriate, or very socially appropriate.*
Second-Stage	*For each of the choices, please indicate your guess for the most frequently given by the evaluators. Remember that you will earn money ($2) if your response to a randomly-selected question is the same as the most common rating provided by the evaluators.*
Non-Incentivized Second-Stage	*For each of the choices, please indicate your guess for the most frequently given by the evaluators.*
Krupka-Weber	*For each of the choices, please indicate whether you believe choosing that option is very socially inappropriate, somewhat socially inappropriate, somewhat socially appropriate, or very socially appropriate. To indicate your response, please place a checkmark in the corresponding box. Remember that you will earn money ($2) if your response to a randomly-selected question is the same as the most common response provided in this study.*
Non-Incentivized Krupka-Weber	*For each of the choices, please indicate your guess for the most frequently given rating in this study.*

A.2 Effect of incentivization on rating variance

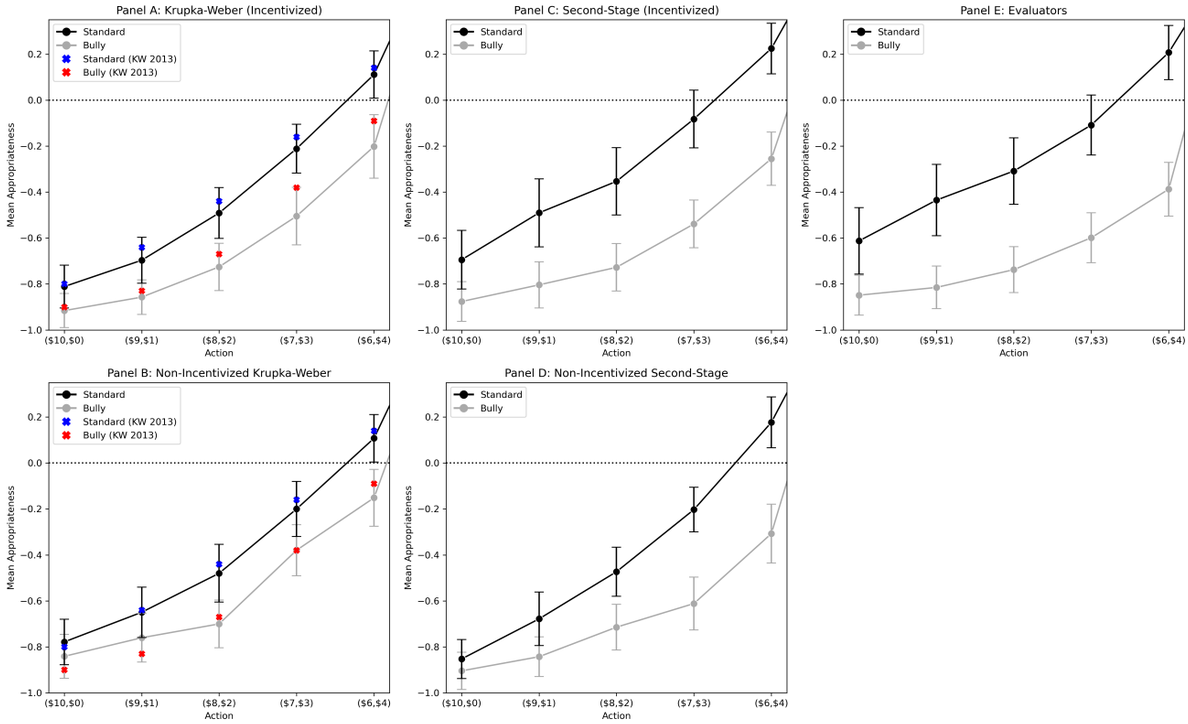
Table A2: Effects of Incentivization on Rating Variance

	Non-Incentivized			Incentivized			
Outcome	Obs	Mean	Std Dev	Obs	Mean	Std Dev	F-value
S (10,0)	228	-0.752	0.498	162	-0.759	0.502	0.983
S (9,1)	228	-0.593	0.572	162	-0.604	0.568	1.013
S (8,2)	228	-0.424	0.561	162	-0.429	0.584	0.924
S (7,3)	228	-0.172	0.515	162	-0.153	0.531	0.941
S (6,4)	228	0.162	0.490	162	0.162	0.489	1.005
S (5,5)	228	0.794	0.362	162	0.828	0.376	0.925
S (4,6)	228	0.441	0.558	162	0.513	0.553	1.018
S (3,7)	228	0.309	0.613	162	0.287	0.665	0.847
S (2,8)	228	0.146	0.736	162	0.157	0.762	0.935
S (1,9)	228	0.073	0.791	162	0.099	0.821	0.929
S (0,10)	228	0.051	0.842	162	0.044	0.879	0.918
B (10,0)	218	-0.864	0.381	143	-0.894	0.355	1.155
B (9,1)	218	-0.807	0.413	143	-0.827	0.397	1.080
B (8,2)	218	-0.718	0.439	143	-0.727	0.447	0.964
B (7,3)	218	-0.534	0.494	143	-0.524	0.487	1.028
B (6,4)	218	-0.288	0.541	143	-0.231	0.541	0.997
B (5,5)	218	0.817	0.419	143	0.808	0.412	1.035
B (4,6)	218	0.438	0.588	143	0.497	0.509	1.338*
B (3,7)	218	0.289	0.656	143	0.360	0.608	1.161
B (2,8)	218	0.179	0.715	143	0.281	0.692	1.067
B (1,9)	218	0.099	0.764	143	0.238	0.751	1.021
B (0,10)	218	0.062	0.829	143	0.178	0.817	1.028

Notes: This table tests the statistical difference between the variance of ratings of non-incentivized participants and incentivized participants. Outcomes are presented with the moving player first such that the outcome (10,0) indicates the dictator ending up with $10 and the other player recieving $0. The "S" and "B" before the outcome represent the Standard and Bully variants of the dictator game, respectively.

A.3 Analysis by experimental condition

Figure A1: Giving and Taking Norms by experimental condition



Notes: This figure shows the mean appropriateness ratings of each action in the standard and bully environments of the dictator game for outcomes ($10,$0) to ($6,$4). Each panel corresponds to an experimental condition. For both incentivized and non-incentivized versions of the Krupka-Weber elicitation method, the corresponding ratings from the original Krupka & Weber (2013) study are shown. Error bars reflect 95 percent confidence intervals.