Hande Erkut*

Incentivized Measurement of Social Norms Using Coordination Games

https://doi.org/10.1515/auk-2020-0004

Abstract: Social norms are important determinants of behavior. Hence, we need reliable methods to identify them in order to increase the predictive and explanatory power of models that aim to predict human behavior. In this paper, I will focus on a norm measurement method proposed by Krupka and Weber. In particular, I will discuss whether social norms elicited using this method are malleable, and whether these norms are good predictors of behavior.

Keywords: social norms, norm measurement, incentives

1 Introduction

Social norms have long been recognized by social scientists as being important social constructs governing behavior. Social norms are collectively perceived rules that prescribe or proscribe actions, and behavior such as cooperation, reciprocity, and retribution can potentially be explained by such rules. Nonetheless, in principle any action can be post-hoc explained by following some norm, which makes it harder to refute norms-based explanations (Fehr/Schurtenberger 2018). This is the reason why it is crucial to find reliable ways to identify social norms empirically.

There are direct and indirect ways to identify social norms. The indirect way of identifying the existence of a social norm is by using punishment mechanisms, and a number of studies have used this method (e.g., Ostrom et al. 1992; Fehr et al. 2002; Bicchieri et al. 2011). In a typical economic game with a punishment option, players can choose to decrease another player's payoffs by paying a small cost. It is assumed that the players punish socially inappropriate behavior, hence what a person ought not do in such a game is identified by observing the punished actions. There are two problems with this method. First, punishment might also be motivated by factors other than socially inappropriate behavior. Second, not everybody might want to engage in costly punishment, and with this method we only elicit the norms of those people who like to punish others.

***Corresponding author: Hande Erkut,** WZB Berlin Social Science Center, e-mail: hande.erkut@wzb.eu

The straightforward, direct method to measure social norms is by asking people to rate how socially appropriate an action is in a given situation. The problem with this method is that the people who answer these questions do not have an incentive to report their normative beliefs truthfully. To address this problem, Krupka and Weber (2013) (henceforth KW) developed a method in which people have an incentive to report their true normative beliefs.[1] In this paper, I will first explain the KW method for measuring social norms and review how well the norms measured with this method predict behavior. Second, I will discuss several criticisms of the method. Finally, I will conclude.

2 The Krupka & Weber Method

2.1 Norm Measurement

The KW method relies on two important characteristics of social norms, following the definition of Elster (1989). First, social norms are based on actions rather than outcomes. Hence, social appropriateness is defined by the action that results in a particular outcome, and it is not defined by the outcome itself. Following this definition, different actions that lead to the same outcome can have different appropriateness ratings. Second, social norms are rules that are collectively perceived by the members of the society.

According to the utility framework KW uses, a decision-maker cares about two things when taking an action: the monetary payoff an action brings,[2] and the action's perceived social appropriateness. In order to measure the social appropriateness of an action, KW presents the set of all possible actions in a given situation and asks respondents to rate the social appropriateness of each action as 'very socially inappropriate', 'somewhat socially inappropriate', 'somewhat socially appropriate', or 'very socially appropriate'. Respondents earn a monetary reward if they can successfully coordinate on the normative beliefs of the society. In particular, if the appropriateness rating they indicate for a randomly selected action is the appropriateness rating selected by the most participants in their experimental session, they earn a monetary reward. Hence, they are incentivized to

1 Other studies that use incentivized methods to elicit social norms include Bicchieri/Xiao 2009 and Bicchieri/Chavez 2010.

2 In this utility framework, the monetary payoff an action brings represents the material consequence of that action. In principle, the material consequences do not always have to be monetary. For instance, the material consequences of an action can also be food, gifts or even pain (See Erkut 2018 for social norms elicitation in the pain domain).

state social norms and not their personal opinions about what is the right thing to do. With this method, KW ensures that two important characteristics of social norms—action-based and collectively perceived—are reflected in the respondents' ratings.

By using two variants of dictator games, KW shows how different actions leading to the same outcomes can have different social appropriateness ratings. In both dictator games, there are two players where one of them is the decision-maker and the other one is the passive player. In the standard dictator game, the decision-maker receives $10 and the passive player receives nothing. The decision-maker has to decide how much of the $10 to give to the passive player in $1 increments. The amount he gives has to be between $0 and $10. In the bully dictator game, both the decision-maker and the passive player receive $5, and the decision-maker has to decide whether to give money to or take money from the passive player. The amount given or taken has to be between $0 and $5. Both dictator games have the same 11 possible outcomes, but the decision-maker has to take different actions in each game to obtain these outcomes. For instance, in order to obtain the outcome where the decision-maker ends up with $6 and the passive player ends up with $4, the decision-maker has to give $4 to the passive player in the standard game and she has to take $1 from the passive player in the bully game.

KW presented the above games to experimental subjects (each subject was presented with one variant of the game), and asked them to rate the social appropriateness of each action a decision-maker could take in the game. Note that these subjects did not play the game, they only rated the social appropriateness of the actions in the game. Subjects earned money if their stated appropriateness rating of a randomly chosen action matched with the modal appropriateness rating for this action in their experimental session. So, for instance, suppose that a subject takes part in an experimental session where she rates the appropriateness ratings of the actions in a standard dictator game. Moreover, suppose that giving $5 is randomly chosen as the action that will determine the payment in this experimental session. If the most commonly chosen appropriateness rating for giving $5 is 'very socially appropriate' in the session, and if the subject also rates giving $5 as very socially appropriate, she receives a payment. If she gives a different appropriateness rating for the specified action, she does not receive a payment.

2.2 How Well Do the Measured Norms Using the KW Method Predict Behavior?

KW hypothesizes that taking money would be seen as more socially inappropriate than giving money even in cases where it produces the same outcome. To illustrate, the expectation is that taking $2 would have a lower appropriateness rating than giving $3, although the decision-maker ends up with $7 and the passive player ends up with $3 in both cases. The results of their experiment are in line with their expectations. Given an outcome level, people evaluate actions involving taking money as less socially appropriate than the actions that involve giving money.

In order to examine whether the measured norms predict behavior in the standard and the bully dictator game, KW ran an experiment with a different set of people, and let them play one of those games.[3] To investigate whether the likelihood of choosing a particular action in those games is influenced by the monetary payoff the action brings and by the average social appropriateness rating of this action,[4] they ran a conditional (fixed-effects) logistic regression. Note that the monetary payoff of an action leading to the same outcome is identical in the standard and bully dictator games, whereas the average appropriateness ratings are not. For instance, the actions leading to the outcome of $7 for decision-maker and $3 for the passive player have the same monetary payoff—which is $7—for decision-maker in both the standard and bully dictator games. However, the action leading to this outcome has a different (and lower) average social appropriateness rating score in the bully game than in the standard game. The results suggest that the likelihood of choosing an action increases with both the monetary payoff and the average social appropriateness rating score.

KW further measured the norms for different variants of the dictator game (dictator game with and without an opt-out option (Lazear et al. 2012), dictator game with a taking option (List 2007), and dictator game with hidden information (Dana et al. 2006)), and investigated whether the norms for these games could explain the behavior. These dictator games share the common element of producing results that cannot be explained by standard social preference models. KW's re-

3 As the subjects who state social norms and who play the games come from the same subject pool, they are assumed to share the same social norms.

4 In order to calculate the average social appropriateness rating, the social appropriateness ratings given by people in the previous experiment are converted to the following numerical scores: very socially inappropriate is -1, somewhat socially inappropriate is -0.33, somewhat socially appropriate is 0.33, and very socially appropriate is 1.

sults suggest that the measured norms for these games can successfully account for the behavior observed.[5]

There is also evidence to suggest that social norms are not always the best predictor of behavior. For instance, Gächter et al. (2013) show that a social preferences model assuming inequality-averse preferences more accurately predicts peer effects in a three-person gift exchange game than social norms. Also, Krupka et al. (2017) suggest that a guilt aversion or a lying aversion model combined with social norms better explains the behavioral differences in double dictator games and Bertrand games with and without informal agreements than social norms do alone.

3 Comments on the Method

In this section I will first discuss the KW method for identifying social norms. In particular, I will investigate whether people need to be incentivized to elicit social norms, and whether the social norms elicited using the KW method are malleable. Second, I will discuss ways to improve the predictive power of the measured norms.

3.1 Eliciting Social Norms

KW found a novel way to elicit people's shared beliefs of what a person ought to do in an incentivized way, but the question is whether people actually need to be incentivized to state their social norms at all. It might be enough to ask them the socially appropriate action in a given situation without providing monetary incentives. In fact, Veselý (2015) investigated whether incentives matter when eliciting social appropriateness ratings using the KW method in an ultimatum game, and found no difference between the incentivized and non-incentivized ratings. Yet, as in dictator games, the right way to act is fairly straightforward in ultimatum games, so the social and personal norms—what one personally believes is the right thing to do—are more likely to align.[6] In cases where these two do not align, the lack of incentives may prevent researchers from eliciting the correct beliefs.

5 The KW method for measuring norms is also used in other contexts such as peer effects (Gächter et al. 2013), bargaining (Banerjee 2016), discrimination (Barr et al. 2018), trust (Krupka et al. 2017), distributing harm (Erkut 2018) and antisocial behavior (Behnk et al. 2019).

6 Personal norms are different from social norms in the sense that social norms are rules that are collectively perceived by the society, whereas personal norms are not.

In the absence of incentives, people will be more inclined to state their personal beliefs about what is the right thing to do due to false consensus bias or due to the desire to share the same ethical point of view with society. Incentives can help to overcome these biases and help to elicit beliefs as accurately as possible by making sure that reporting incorrect beliefs has a cost. Without such a cost, people may be more likely to state their personal norms than the social norms in cases where the former is different from the latter.

Incentives become even more crucial when social norms are elicited from people who previously made the decisions in the situation that they are evaluating. Rustichini and Villeval (2014) elicited non-incentivized personal fairness judgments from people before and after they played dictator, ultimatum, and trust games. In particular, they asked subjects about the interval of the fair and unfair actions in those games. One week later, they asked people to play these games and again asked the interval of the fair and unfair actions.[7] They found that people adjust their fairness evaluations by including their actions in the 'fair actions' interval. Although this study focuses on personal fairness judgments and not on social fairness judgments, it shows us how judgments can be prone to malleability, i.e., they can be manipulated to justify actions, in the absence of incentives.

When incentives are used, the elicited social norms are not malleable. Erkut et al. (2015) compared the social norms elicited using the incentivized KW method from the spectators with the norms elicited from the people who had played the dictator game either as a decision-maker or as a passive player (stakeholders). Their results suggest that the elicited norms do not differ between spectators and stakeholders. Moreover, the subjects who were put in the role of decision-makers and passive players reported similar social norms even though they were incentivized to coordinate with the people in their own role while reporting social norms. Hence, social norms elicited using the KW method are not malleable in dictator games.

Yet, as shown by KW, the most socially appropriate action to take is straightforward in standard dictator games—50-50 split—, and this result may not hold in more complex games with multiple social norms. In contexts with multiple norms, people may be more inclined to report norms that are in favor of their role and interest in the game. For instance, consider the following dictator game with a production stage presented in Cappelen et al. (2007). In the first stage, the decision-maker and the passive player do a real-effort task to produce the amount of pie to be divided, and in the second stage decision-maker divides the produced pie

7 The subjects did not know that they would be playing these games for real when they reported their fairness judgments as spectators.

between herself and the passive player. In this game, an egalitarian fairness norm requires equal division of the pie independent of the effort exerted in the production stage. On the other hand, a libertarian fairness norm requires the division of pie in proportion to the effort exerted. If the social norms of this game are elicited from those who played the game, people may report the fairness norm that is in their best interest. So, for instance, a player who exerted greater effort potentially reports the action that is in line with the libertarian fairness norm as appropriate. Hence, in contexts with multiple norms, using the KW method to elicit social norms from the people who played the game previously may not be appropriate.

3.2 Following Social Norms

Bicchieri (2006) defines the conditions for the existence of a social norm and the conditions for following a norm as follows: For a social norm to exist, people in a population should know that there is a rule applying to a situation. For a person to conform to the rule, she should expect that a sufficiently large subset of the population also conforms to the rule (empirical expectations), and she should expect that others expect her to conform to the rule and/or she may be punished for not following the rule (normative expectations).

The KW method elicits social norms and normative expectations by asking people about the social appropriateness of the actions that could be taken, and by incentivizing them to coordinate on the social appropriateness ratings of each action. Yet it does not give us information on empirical expectations. Hence, it does not fully identify the conditions to follow the norm. Identifying the conditions to follow a norm is important for predicting norm-following behavior and for understanding the reasons behind the heterogeneity of the norm-following behavior among people. The fact that empirical expectations have not often been elicited and incorporated with the social norms might be the reason why the elicited social norms is not the best predictor of behavior in the aforementioned cases of peer effects and dictator games with informal agreement. For instance, as previously discussed, Krupka et al. (2017) found that social norms incorporated with a guilt aversion model—which assumes that people get a feeling of guilt if they do not live up to others' expectations—is a better predictor of behavior than social norms alone. This finding gives us a hint on how combining second-order empirical expectations—what I believe about others expect me to do—and social norms ameliorates predictive power.

Other potentially important determinants for following a norm is agreement with the social norm as pointed out by Erkut and Reuben (2019), and rule-following behavior (Kimbrough/Vostroknutov 2016). A person who agrees

with the social norm and who further internalizes the norm as the personal norm is potentially more likely to follow the norm than a person who believes that the norm is not sensible. Also, a person who is more likely to follow rules in general is potentially more likely to follow social norms. In fact, Kimbrough and Vostroknutov (2016) elicited people's rule-following preferences using a novel method and showed that social norms are a greater determinant of the behavior of people who are more likely to follow rules.

4 Conclusion

The above discussion on the identification of social norms using the KW method, and whether the elicited norms predict behavior, provides us with two main insights. First, incentivizing the elicitation of social norms is necessary to ensure that elicited norms are not malleable, but it might not always be sufficient. When the norms are elicited from spectators, incentivizing people to coordinate on the shared belief on what is the right thing to do makes sure that subjects have an incentive to report social and not personal norms. When the norms are elicited from stakeholders, incentivizing becomes even more important. In order to avoid contradicting themselves, subjects who play the game for real will be inclined to state the actions they took in the game as being more appropriate than the other available actions. Incentivizing them to report correct beliefs on the appropriate action makes it costly to report other actions as appropriate. Nevertheless, norms elicited using the KW method may still be malleable in situations with multiple norms, if the method is used to elicit the social norms of people who were previously decision-makers in the situations evaluated. In multiplicity of social norms, people who have different roles and interests in the situation might report the norms that serve their self-interest. Hence, eliciting social norms from the people who previously played the game may not be the best strategy in contexts with multiple norms.

Second, although the measured norms using the KW method are helpful for predicting behavior, additional belief and preference measures can be utilized to predict norm-following behavior more precisely and to account for the reasons for heterogeneity in the norm-following behavior. In particular, social norms elicited using the KW method may be combined with empirical expectations, rule-following preferences, and personal norms.

As a final note, although the elicited social norms using the KW method can predict behavior in a variety of contexts, we cannot claim that the elicited norms cause behavior in those contexts. For instance, KW showed that both social norms

and behavior are different in the standard and bully dictator games, and that different social norms can predict behavioral differences in those games. Nevertheless, it is hard to be certain that social norms are the only things that change between these two games, which makes it harder to infer a causal relationship between social norms and behavior.

References

Banerjee, R. (2016), On the Interpretation of Bribery in a Laboratory Corruption Game: Moral Frames and Social Norms, in: *Experimental Economics* 19, 240–267

Barr, A./T. Lane/D. Nosenzo (2018), On the Social Inappropriateness of Discrimination, in: *Journal of Public Economics* 164, 153–164

Behnk, S./L. Hao/E. Reuben (2019), *Shifting Normative Views: On Why Groups Behave More Antisocially Than Individuals*

Bicchieri, C. (2006), *The Grammar of Society: The Nature and Dynamics of Social Norms*, Cambridge

—/A. Chavez (2010), Behaving as Expected: Public Information and Fairness Norms, in: *Journal of Behavioral Decision Making* 23, 161–178

—/E. Xiao (2009), Do the Right Thing: But Only If Others Do So, in: *Journal of Behavioral Decision Making* 22, 191–208

—/E. Xiao/R. Muldoon (2011), Trustworthiness Is a Social Norm, But Trusting Is Not, in: *Politics, Philosophy & Economics* 10, 170–187

Cappelen, A. W./A. D. Hole/E. Ø. Sørensen/B. Tungodden (2007), The Pluralism of Fairness Ideals: An Experimental Approach, in: *American Economic Review* 97, 818–827

Dana, J./D. M. Cain/R. M. Dawes (2006), What You Don't Know Won't Hurt Me: Costly (But Quiet) Exit in Dictator Games, in: *Organizational Behavior and Human Decision Processes* 100, 193–201

Elster, J. (1989), Social Norms and Economic Theory, in: *Journal of Economic Perspective* 3, 99–117

Erkut, H. (2018), *Social Norms and Preferences for Generosity are Domain Dependent*, WZB Discussion Paper

—/D. Nosenzo/M. Sefton (2015), Identifying Social Norms Using Coordination Games: Spectators vs. Stakeholders, in: *Economics Letters* 130, 28–31

—/E. Reuben (2019), Preference Measurement and Manipulation in Experimental Economics, in: *Handbook of Research Methods and Applications in Experimental Economics*, Cheltenham

Fehr, E./U. Fischbacher/S. Gächter (2002), Strong Reciprocity, Human Cooperation, and the Enforcement of Social Norms, in: *Human Nature* 13, 1–25

—/I. Schurtenberger (2018), Normative Foundations of Human Cooperation, in: *Nature Human Behaviour* 2, 458–468

Gächter, S./D. Nosenzo/M. Sefton (2013), Peer Effects in Pro-social Behavior: Social Norms or Social Preferences?, in: *Journal of the European Economic Association* 11, 548–573

Kimbrough, E. O./A. Vostroknutov (2016), Norms Make Preferences Social, in: *Journal of the European Economic Association* 14, 608–638

Krupka, E. L./S. Leider/M. Jiang (2017), A Meeting of the Minds: Informal Agreements and Social Norms, in: *Management Science* 63, 1708–1729

—/R. A. Weber (2013), Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?, in: *Journal of the European Economic Association* 11, 495–524

Lazear, E. P./U. Malmendier/R. A. Weber (2012), Sorting in Experiments with Application to Social Preferences, in: *American Economic Journal: Applied Economics* 4, 136–163

List, J. A. (2007), On the Interpretation of Giving in Dictator Games, in: *Journal of Political Economy* 115, 482–493

Ostrom, E./J. Walker/R. Gardner (1992), Covenants with and without a Sword: Self-Governance Is Possible, in: *The American Political Science* 86, 404–417

Rustichini, A./M. C. Villeval (2014), Moral Hypocrisy, Power and Social Preferences, in: *Journal of Economic Behavior & Organization* 107, 10–24

Veselý, S. (2015), Elicitation of Normative and Fairness Judgments: Do Incentives Matter?, in: *Judgment and Decision Making* 10, 191