Luise Görges and Daniele Nosenzo*

Measuring Social Norms in Economics: Why It Is Important and How It Is Done

https://doi.org/10.1515/auk-2020-0012

Abstract: Experimental economics offers new tools for the measurement of social norms. In this article, we argue that these advances have the potential to promote our understanding of human behavior in fundamental ways, by expanding our knowledge beyond what we learn by simply observing human behavior. We highlight how these advancements can inform not only economic and social theory, but also policymaking. We then describe and critically assess three approaches used in economics to measure social norms. We conclude our overview with a list of recommendations to help empirical researchers choose among the different tools, depending on the nature and constraints of their research projects.

Keywords: social norms, injunctive norms, second-order beliefs, social appropriateness, measuring norms, experimental economics

1 Introduction

In recent years economists have grown increasingly interested in understanding the influence on economic behavior of '*social norms*', shared understandings within a social group about what is considered acceptable or unacceptable behavior in a given situation. A large number of economic experiments have produced evidence suggesting that social norms exert indeed a powerful influence on human decision-making in a vast array of social and economic situations. Norms, for instance, have been found to play an important role for cooperative behavior (Reuben/Riedl, 2013), fair sharing (Krupka/Weber 2013; Gächter et al. 2017), honesty (Abeler et al. 2019), female labor force participation (Görges 2020), discrimination (Barr et al. 2018), and corruption (Gneezy et al. 2019).

The focus on norms as a key explanation for human behavior represents an important paradigm shift in economics, as it moves the discipline away from an

Luise Görges, Institute of Economics, Leuphana University Lüneburg, Lüneburg, Germany, e-mail: luise.goerges@leuphana.de
***Corresponding author: Daniele Nosenzo,** Department of Economics and Business Economics, Aarhus University, Aarhus, Denmark, e-mail: daniele.nosenzo@econ.au.dl

approach dominated by methodological individualism by explicitly incorporating social forces among the key determinants of human action. As discussed in greater detail in Kliemt (2020) and Vostroknutov (2020), this implies abandoning the notion that individual behavior can be rationalized simply by referring to self-interested considerations, or to individualistic intrinsic motivations to behave pro-socially, and recognizing instead that a full account of the role of social forces in shaping and interacting with individual behavior is necessary to explain human decision-making. As a result, several theories of human behavior have been formulated that explicitly incorporate social norms to improve the predictive ability of economic models (e.g., Akerlof/Kranton 2000; Bénabou/Tirole 2006; Andreoni/Bernheim 2009).

This paradigm shift that is occurring in economics, however, is not only conceptual, but also *empirical*. If social norms are to become an important component of economic theory, it is crucial to be able to develop robust and accurate empirical tools that allow us to measure them alongside the traditional data sources that we use to study human behavior (data on economic agents' choices, collected via experiments, surveys or observational studies). Measuring norms, and their influence on individual choices, is in fact an essential step toward assessing the extent to which theories of social norms explain and predict social and economic behavior. Thus, during the last decade or so, economists have started to develop a number of *empirical tools to measure social norms*. This article focuses on these developments in the discipline, by arguing why it is important to measure norms directly and by reviewing the empirical methods that have been proposed in the literature to achieve this. [1]

After briefly outlining a conceptual framework to guide our discussion of social norms *section 2*), in the first half of the article we argue how the empirical measurement of norms can produce (and has already produced) important advancements in knowledge that go beyond what we could have learned by simply observing human behavior (*section 3*). In doing so, we highlight how these advancements can inform not only economic and social theory, but also policymaking—because improving our understanding of the social forces that govern human behavior is crucial to evaluate which type of interventions may be more suitable and effective to instigate and achieve behavioral change. In the second half of the paper, we then describe and critically assess three approaches that are used by economists

1 A full comparative, cross-disciplinary review of the topic is beyond the scope of our article. Rather, its purpose is to provide an accessible overview to scholars in other disciplines of the methodological work to empirically measure norms that is currently being undertaken in economics. We hope that this can serve as a springboard for a dialogue between economists and other social scientists about this important topic and lead to fruitful interdisciplinary collaborations.

to measure social norms (*section 4*). Each approach has its advantages and disadvantages, and we therefore conclude our overview with a list of recommendations to help empirical researchers choose among the different tools depending on the nature and constraints of their research projects.

2 Conceptual Framework

Before discussing the *why* and *how* of the empirical measurement of social norms, we begin by sketching a conceptual framework that will form the basis of our discussion. Many of the economic models of human behavior that explicitly incorporate the influence of social forces on individual decision-making build on the observation that individuals care about how others perceive them (e.g., Bernheim 1994; Akerlof/Kranton 2000; Bénabou/Tirole, 2006; Andreoni/Bernheim 2009; Krupka/Weber 2013 Kimbrough/Vostroknutov 2016; d'Adda et al. 2020). Caring for social approval implies that individuals will be concerned with meeting others' expectations about what behavior is approved and disapproved of in their social network. We can think of social norms as informal *rules of conduct that embody beliefs about which actions are approved or disapproved of* in a specific context by a given social group.

Individuals are motivated to follow norms, even when this requires taking actions that run counter the individuals' material self-interest, as doing so gains them the approval (and spares them the disapproval) of their social group. The utility gains from social approval, and the opportunity costs of deviating from self-interest maximization, will be important factors to determine the extent to which individuals comply with social norms in any given situation (for a more detailed discussion, see, e.g., Sugden 1998a; 2000; Bicchieri 2006; 2017; Fehr/Schurtenberger 2018).[2]

There are a few aspects of this framework that are important to clarify before moving to the discussion of the empirical measurement of norms. A first point to notice is that the framework sketched above focuses on what the literature refers to as *'injunctive norms'*, i.e., beliefs about what one *ought to do*, or not do,

2 Although social approval and disapproval are important types of (non-material) social sanctions that individuals can use to enforce norm compliance (for instance, taking the form of conferral of esteem or prestige, stigmatization, avoidance, jeering, etc.), other mechanisms may influence the extent to which individuals comply with norms. For instance, in some cases norm violations may be punished with *material* sanctions (imprisonment, loss of property, violence). When norms are internalized, norm compliance may also be sustained by *internal* sanctions, i.e., sanctions that the individual imposes on herself, for example through feelings of guilt or remorse.

in a particular situation. Injunctive norms are different from *'descriptive norms'*, which instead refer to what is *commonly done* in a situation. In some theories of social norms, descriptive norms also play an important role in shaping individual behavior, as social approval and disapproval can also be conferred to behaviors that blend in or deviate from what is the most commonly occurring behavior in a social group (see, e.g., Bicchieri 2006; 2017; Vostroknutov 2020). However, our discussion about the measurement of norms is focused on injunctive norms, partly because measurement of descriptive norms is more straightforward and partly because it has been used in the literature for a relatively longer time than the measurement of injunctive norms.[3]

Second, the beliefs that support injunctive norms are *second-order beliefs* about what is appropriate or inappropriate in a given context. By second-order beliefs, we mean beliefs that the individual holds about what others believe is appropriate or inappropriate (i.e., beliefs about others' beliefs, hence 'second-order'). These are different from *first-order beliefs* of appropriateness, which instead are beliefs about what the individual *personally* considers appropriate or inappropriate. This distinction is important for two reasons. First, conceptually, the desire for *social* approval requires that the individual forms beliefs about what *others* in their social group approve or disapprove of; their own view of what is appropriate is not necessarily relevant. Second, first- and second-order beliefs of appropriateness may not necessarily coincide. For instance, second-order beliefs may be systematically miscalibrated relative to individuals' underlying first-order beliefs, i.e., individuals may believe others endorse a certain action while no one personally does. This phenomenon is known as *pluralistic ignorance* and will be discussed in detail below.

A final point worth emphasizing is that, while it is common to discuss social norms by referring to specific actions that the norm prescribes or proscribes (e.g., 'give up your seat to an elderly person in a crowded public transport'), our conceptual framework clarifies that norms actually embody beliefs about *entire profiles of actions*, including those that deviate from the prescribed/proscribed action. For instance, to study the social appropriateness of giving up one's seat to the elderly in a crowded public transport, it is equally important to know how much approval one would obtain by giving up one's seat as it is to know how much disapproval one would incur were one *not* to give up the seat. Knowing the approval/disapproval

3 Descriptive norms rely on beliefs about what constitutes common behavior in a given situation. These beliefs (sometimes called *'empirical expectations'*, e.g., Bicchieri 2006; 2017) can be measured by asking individuals their guess about the most prevalent behavior in a decision situation. These guesses can be incentivized by paying individuals based on the accuracy of their predictions.

associated with each behavior is crucial as it allows to reconstruct the *structure of social incentives* that support the norm. This is important because the social incentives for following a norm may differ across social groups or social situations, with real implications for the extent to which the individual will feel pressured to comply with the norm. In the public transport example, not giving up one's seat to an elderly person may be considered extremely inappropriate in some groups, but only mildly condemned in others, generating a stronger social pressure to comply in the former case than in the latter.

Before we go into the details of the tools economists use to measure norms, we will expand on how empirical methods of measuring norms that build on the framework we sketched in this section—norms that prescribe *what ought to be done* and operate through *second-order beliefs* of what is approved/disapproved of in a given situation, and over a *range of behaviors*—can produce important insights that go beyond what we could have learned by simply observing human behavior.

3 Why It Is Important to Measure Norms: Practical Examples

3.1 What We Learn by Measuring (Injunctive) Social Norms

In this section, we will contend the notion that mere documentation of behavioral patterns that are consistent with norm-following provides sufficient evidence of the existence of social norms. Rather, we will argue, it is essential to *actually measure the second-order beliefs* that underlie the norm. Such a measurement of norms is important because it informs both social and economic theories (promoting the advancement of scientific knowledge) and policymaking (promoting the development of better and more effective interventions).

To appreciate the importance of measuring norms, consider gender differences in the labor market as an example. The *gender gap in employment* has been stalling over the past forty years in most industrialized countries of the world, as women's labor force participation in the 1980s discontinued the remarkable growth it had seen until then and has not caught up any further with men's level of participation since (Goldin, 2006; Blau/Kahn, 2013). In recent years, policy makers around the world have expressed considerable frustration over the stubbornness of the gender employment gap and other dimensions of inequality—e.g., the European Institute for Gender Equality in its most recent report (EIGE, 2019) prominently lamented a "snail's pace towards gender equality"—partly because traditional economic explanations would suggest a faster convergence. For instance, the closing of the

gender gap in educational attainment, a dimension on which women are now even overtaking men in some countries (EIGE, 2019), did not result in a closing of the gender employment gap.

So what could explain the persistence of gender differences in employment? Policymakers increasingly point to *gender norms* as a key obstacle to more equal participation of men and women in the labor force (OECD 2017; EIGE 2019; European Commission 2020). For instance, the European Commission identifies gender norms as the "root cause of gender inequality (…)" which "limit (…) choices and freedom, and therefore need to be dismantled" (European Commission 2020). The argument is motivated with figures from the most recent Eurobarometer, which show that 44% of Europeans think that "the most important role of a woman is to take care of her home and family" (European Commission 2017). Taken at face value, this argument implies that one reason why women spend less time in the labor market than men is that society condemns a woman who spends more time at work than at home with her family. That is, there is a *gender norm* that prescribes that women spend time at home with their children rather than at work. Although intuitively appealing, providing hard evidence in support of this argument is surprisingly difficult. Indeed, we will argue that the necessary evidence can only be collected through the *empirical measurement of the gender norm* that policymakers suspect underlies the employment gap.

To clarify this point, we start by sketching a simple model of labor supply. The model is useful because it allows us to illustrate, in a simple way, the possible channels that may lead a woman to supply less labor than a man. This exercise will illuminate how simply observing a certain behavioral pattern that are consistent with the gender norm (here: women supplying less labor than men) is not sufficient to provide evidence that the norm does in fact exist, let alone causes this behavior, because other channels may plausibly produce the same pattern of behavior.

Assume that individual i derives utility from market goods, M, and time at home with her child, C. The amount of M she can buy depends on the time she spends working in the labor market, t_i, which, together with her wage rate, w_i, determines her available income. As a result, M increases in t_i and w_i. The amount of C she consumes is simply given by her remaining available time, $1 - t_i$ (her total time budget is normalized to 1) and is therefore a decreasing function of market time. Finally, we assume that the individual also cares about the approval or disapproval from her peers, which may in turn depend on the extent to which the woman conforms to a gender norm proscribing that a woman spend more time in the labor market than t_G hours or, equivalently in this example, prescribing she spend a minimum amount of time, $1 - t_G$, at home with her child. Consequently, as soon as she violates that norm, N decreases in $t_i - t_G$, i.e., in the difference

between the time a woman actually spends in the labor market and the maximum time she ought to spend, according to the gender norm.

$$U_i(M(w_i, t_i), C(1 - t_i); N(t_i, t_G))$$

The agent maximizes her utility U_i — increasing in the arguments M, C and N — by choosing her time in the labor market, t_i, such as to achieve her optimal levels of consumption of M and C, and to reduce the costs arising from a violation of the gender norm N. This model, albeit simple, highlights three plausible reasons why women may systematically supply less time to the labor market than men. The first reason is that the gender norm prescribes lower levels of market time for women than for men. Since norm violations yield psychological costs, women who are otherwise identical to men spend less time in the market and more time at home. This is the explanation suggested by policymakers. However, two other reasons may spur systematic gender differences in labor supply, even when there is no norm prescribing differential ideal levels of market time for the genders. One is that women's returns to time spent in the labor market may systematically fall short of men's. In our simple model here, this is captured by wage rates, w_i, which may differ systematically across genders, e.g., due to labor market discrimination. Another reason why women supply less time to the labor market than men could be a differential valuation of market goods relative to time spent with children.

This illustrates why simple observation of the persistence of lower female labor force participation in a given society cannot be taken as conclusive evidence of the existence of a gender norm in that society. The same behavioral regularity could also be produced by either of the other two channels. What is more, the three channels may be interlinked, that is, norms may not only exert an effect on behavior through direct, psychological costs that accrue to individuals who experience social disapproval from violating social norms. It is, in fact, likely that norms also affect the other two parameters given which individuals are optimizing, gender differences in wage rates or in the relative valuation of time spent with children. For example, it has been argued that employer discrimination is positively correlated to the pervasiveness of gender differentiating norms (Givati/Troiano 2012). Consequently, to understand whether and how gender norms affect behavior, mere observation of a behavioral regularity (women supplying less time to the market than men) cannot be sufficient. Only by *explicitly measuring* the actual

gender norms that prevail in that market can one exclude (or confirm) that norms are a possible explanation for the market outcome.[4]

Disentangling the influence of social norms on women's labor force participation from that of other factors is not merely an intellectual exercise; it can offer insights that are of high practical value for the design of public policy. To illustrate this, consider a prominent example that is often cited as evidence for the existence of gender norms and their influence on female labor force participation, the case of East Germany. Shortly before reunification in 1989, East German women's participation rate was above 90%, nearly identical to that of East German men. By contrast, female labor force participation in West Germany did not exceed 60% and was almost 25 percentage points lower than the male rate (Beblo/Görges 2018). Until today, more than 30 years after the reunification, differences in the employment gaps persist. East German women, mothers in particular, continue to be active in the labor market at rates that are more similar to German men than West German women do.

It is often asserted that the persistent differences in female labor supply between East and West Germany are evidence for the existence and lasting influence of gender norms, which were reputedly more egalitarian under the socialist rule in the (East German) former GDR. Yet, it is by no means clear that differences in labor supply are due to different gender norms prevailing in East and West Germany, since the separation experience could plausibly have affected other parameters, like the ones in our model, too. For example, employers may be less inclined to discriminate against East German women, perhaps because decades of experience with East German mothers as workers has reduced employers' uncertainty regarding their commitment to paid work. Lower levels of discrimination imply higher returns to time spent in paid work, which could lead East German women to spend relatively more time to the labor market. Similarly, the legacy of a dense net of daycare facilities that was built in the East during separation continues to feed into regional differences in the supply of public childcare until today. This greater availability of high-quality substitutes for women's own time with children in the East may reduce the value of time spent in the home, leading East German women to supply relatively more time to the market.

At the same time, regional differences in employer discrimination or childcare availability may well coexist with, or exist as a result of, regional differences in gender norms. Therefore, even armed with empirical evidence of such regional

4 Note that, to infer a causal effect of norms on behavior, measuring norms is a necessary, but not a sufficient condition. To adequately address potential endogeneity issues and quantify the effect of a change in norms on behavior, one would need to observe an exogenous shift in norms.

differences in discrimination and childcare, one cannot understand whether norms exist and, if they do, whether and how they exert an effect on gender differences in labor supply. That is to say, without explicitly measuring norms, one cannot infer their role in shaping behavior. Yet, how can a policymaker choose a strategy to raise maternal employment without that knowledge? Should she raise the level of publicaly available childcare spaces? Should she mandate that employers follow a gender quota in hiring? Or will both of these measures prove ineffective because women fear to be condemned by their peers for spending 'too much' time in the market?

If norms do play a meaningful role in the labor supply choice of women, another important question that can only be answered by explicitly measuring norms is: who supports them? Employers? Childcare providers? Are women seeking approval for their behavior from other women or men? From colleagues? From women their age (other mothers) or from older women (their own or their partner's mothers)? In case of the figures from the Euro-barometer cited earlier, this boils down to asking: Who are the 44% of Europeans that think women should primarily be taking care of her home and family? Is their approval relevant to women making the decision? The more we know about who supports the norms that drive behavior, the more targeted any intervention can be.

The example shows the importance of measuring norms to build more accurate models of economic behavior. It highlights that, for a policymaker aiming to develop effective strategies towards a certain goal, it is crucial to understand whether, and in what way, social norms present an obstacle towards that goal. In order to gain this understanding, we have argued here that it is rarely sufficient to study behavioral regularities; rather, one must study injunctive norms directly. In the following subsections, we present arguments as to why it is important to measure injunctive norms through *second-order beliefs*, to distinguish personal norms from social norms, and over a *range of behaviors*, to understand the structure of social incentives and how deviations from the norm are evaluated. Our argument relies on examples from recent work in economics that underscore this view.

3.2 Why Measuring Second-order Beliefs is Important

While economists and other social scientists have generally understood the importance of measuring social norms, many approaches rely on imperfect measures, e.g., proxies derived from behavioral regularities (sometimes referred to as descriptive norms) or first-order beliefs. To illustrate how the use of first-order beliefs, i.e., personal opinion about what is appropriate, complicates the empirical challenge of isolating the influence of gender norms on the labor supply decision, consider a

standard item that may at first glance seem well-suited to proxy gender norms. In many surveys around the world, respondents have been asked to state their agreement with the statement: 'A pre-school child suffers when the mother is working' (included, e.g., in multiple waves of the World Value Survey).

A researcher who is interested in studying the impact of gender norms on female labor supply might aggregate agreement with this statement at the society level to assess whether higher levels of agreement in a society are associated with lower levels of female labor supply. Yet, such a relation does not tell us whether women supply fewer hours to the market than men because they feel pressured to conform to societal norms and seek the approval of others. Indeed, the fact that many individuals hold the first-order belief that small children suffer if their mothers work might reveal a higher valuation of time spent with children among women relative to men and, perhaps, a motivation for employers to statistically discriminate against women. In our example of East and West Germany above, systematic differences in the belief that a pre-school child suffers when the mother is working could simply reflect differences in the availability of high-quality public childcare. Consequently, an approach that relies on measuring social norms through aggregation of first-order beliefs likely creates a measure that confounds norms and other relevant determinants of female labor supply.[5]

In some cases, there may even be misalignment between *first- and second-order beliefs*, which may constitute an additional challenge for policy-makers trying to instigate behavioral change. An example of such a case can be found in recent work by Bursztyn et al. (2020), which shows that measuring both first- and second-order beliefs can be desirable in some contexts. The authors study the norms surrounding female labor force participation in Saudi Arabia, where women typically need the permission of their male 'guardian' (usually father or husband) to work outside the household. To study the existence of such a normative principle among the guardians and to assess its effects on female labor force participation, Bursztyn et al. (2020) conduct a norm-elicitation experiment with young Saudi husbands. By eliciting their first- and second-order normative beliefs about the appropriateness of women working outside the home, they show that a large majority of the men in their sample underestimates the support for female labor force participation among other young Saudi men.

The wedge between first- and second-order beliefs documented by Bursztyn et al. (2020) is an example of *pluralistic ignorance*. Importantly for our point here, it

5 This does not mean, however, that first-order beliefs have no influence on behavior. In fact, in a recent paper, Bašić/Verrina 2020 show that both second-order and first-order beliefs exert independent influence on behavior.

shows that by studying first-order beliefs on their own, one would have concluded that in Saudi Arabia there is a norm *favoring* female labor force participation, given that nearly 90% of men find it appropriate for women to work outside the home. This conclusion, however, would have been at odds with the fact that female labor force participation remains very low in the country. The elicitation of second-order beliefs, instead, reveals a widely-held belief among men that society generally condemns female labor force participation and can therefore explain why female labor force participation in Saudi Arabia is so low: male guardians may be reluctant to give their daughters or wives approval to work outside the home, assuming that, by doing so, they would break the prevailing social norm in their country.

The example of Bursztyn et al. (2020) also illustrates the practical usefulness of uncovering a misalignment of first- and second-order beliefs, and the presence of pluralistic ignorance, from a policy perspective. The authors show that randomly correcting husbands' second-order beliefs about the approval rates regarding women's work outside the home among similar men actually increases husbands' willingness to help their wives search for and wives' likelihood to have applied and interviewed for a job outside the home four months later.

3.3 Why Measuring Norms Over a Complete Profile of Actions is Important

The final example we discuss in this section is taken from Krupka/Weber (2013), and illustrates the usefulness of eliciting norms over *ranges of behavior* rather than single actions. Krupka and Weber conduct a series of experiments using two variants of the dictator game (Forsythe et al. 1994). In the standard version of the game, one player (the dictator) is endowed with a sum of money (say, $10). The dictator's task is to divide the $10 in any way she likes between herself and another player (the recipient). The recipient is a passive player and must accept any amount of money that the dictator allocates to him. Krupka/Weber (2013) elicit norms of giving in the standard dictator game, by asking respondents to evaluate the *appropriateness of each of the 11 possible integer divisions* of money between the dictator and the recipient. They find that the most appropriate action is to split the $10 equally. Actions that leave the recipient with less money than the dictator are viewed as inappropriate, and the action that assigns the whole $10 to the dictator is considered the most inappropriate. Actions that leave the recipient with more money than the dictator are judged to be somewhat appropriate, although there is not really a clear consensus among respondents about how to rate these actions.

Krupka/Weber (2013) also conduct a second version of the dictator game (called the 'bully game'). In this version, both the dictator and the recipient are given $5 at

the beginning of the game. The dictator can now give to the recipient any amount of money between $0 and $5, or she can *take* between $0 and $5 from the recipient.[6] As in the standard game, the authors elicit norms of giving (and taking) in the bully version of the game and find that splitting the $10 equally between dictator and recipient is the most appropriate action. Yet, in the bully version, actions that leave the recipient with between $1 and $4 are rated as significantly more inappropriate compared to the analogous actions in the standard version of the game, indicating that the social sanctions against giving little to the recipient are much harsher than in the standard game, plausibly because these actions involve *taking* rather than *giving* money to the recipient.[7] In line with these differences in normative judgments, when a different group of respondents is asked to actually play either version of the game, Krupka and Weber find sizeable behavioral differences across the two games.

Had Krupka and Weber only asked respondents to report what they thought was the *most* appropriate behavior in the game (or only asked them to evaluate the appropriateness of the equal split), they would have concluded that the same norm prevails in the standard and bully versions of the dictator game.[8] In both games, the most appropriate behavior is to give $5 to the recipient. In the bully game, however, giving less than $5 is much worse than in the standard game—i.e., the norm of equality is supported by much stronger social incentives in this version of the game compared to standard. This illustrates the power of eliciting appropriateness judgments about whole ranges of actions available to a decision-maker in a situation, and not only about the action that a researcher may ex-ante think is the one that the norm prescribes or proscribes.

In the next section, we will review and critically assess three empirical elicitation procedures that have been used in the literature to measure norms in accordance with the framework sketched above.

6 Note that the bully game is just a differently-framed dictator game, where the amount to be divided between the players is initially split between them rather than concentrated in the hands of the dictator. However, the possible monetary outcomes of the game (how the money is split between dictator and recipient) are actually identical in the two versions of the game.

7 In the bully game actions that leave the recipient with $1, $2, $3 or $4 imply *taking* money from the recipient, whereas in the standard dictator game these same outcomes are achieved by *giving* money to the recipient.

8 Similarly, in both games the most inappropriate action is to leave the recipient with $0.

4 How Economists Measure Norms: Experimental and Survey Approaches

The discussion in the previous section highlights the importance of developing techniques to measure *injunctive norms*, through *second-order beliefs* about the appropriateness or inappropriateness of *whole ranges of behaviors* that are available to a decision-maker in a specific situation. In the last decade or so, economists have begun to do so. In the remainder of the section, we will describe three different elicitation procedures that have been proposed in the experimental economics literature.[9] In each case, we will critically discuss advantages and disadvantages of the procedures. The section will conclude with a few general considerations about the state of the art in the economics literature and some recommendations for empirical researchers interested in measuring social norms empirically.

4.1 The 'Belief Survey' Method

A straightforward approach to measure social norms is to ask respondents in a survey to report their beliefs about how 'most other people' would rate the appropriateness of various actions available to an agent in a hypothetical decision situation. Asking about *others'* opinion is important because it focuses attention on second-order beliefs rather than on respondents' own opinion about what is appropriate or inappropriate. The focus on the opinion of *most* other people is also important because norms are commonly *shared* understandings of what constitutes appropriate behavior in a given situation. We call this elicitation procedure the *'belief survey'* method, since respondents are simply asked to report their beliefs about other's opinion but do not receive any monetary incentives to guess correctly.[10]

To illustrate the method, imagine that a researcher is interested in studying norms about behaviors related to the diffusion of the recent COVID-19 pandemic,

9 See Görges/Nosenzo (2020) for a discussion of applications of some of these procedures in a labor market context.

10 Several variants of the method have been used in the literature. In some cases, respondents are asked to report their belief about the number of other people who find a certain behavior appropriate or inappropriate. In other cases, respondents are asked to report their belief about how most other people would rate the appropriateness of a certain action. More often, however, this type of uninicentivized questions are used to elicit first-order beliefs instead of second-order beliefs of appropriateness.

such as for instance the practice of wearing a surgical mask in public transport. To elicit this norm, the researcher could include in a survey a short vignette that describes a person waiting for a bus to go to work when realizing she did not bring a surgical mask with her. Respondents could be asked to indicate their beliefs about how most other people would rate the appropriateness of various behaviors available to the person in the vignette, such as, e.g., (i) taking the bus anyway without wearing a mask, (ii) taking the bus covering her mouth and nose with a scarf instead of the mask, (iii) going back home to get a mask and taking the next bus, etc. Respondents could use a Likert scale to indicate how appropriate they believe most other people would find each of the behaviors described in the survey.

Responses to these questions reveal respondents' perception of the social norm surrounding the use of masks in public transport. Importantly, the questions would give the researcher an indication not only about what is the most socially appropriate action (in the example, presumably 'going back home to get the mask'), but also about the relative inappropriateness of different deviations from the most acceptable behavior (e.g., the inappropriateness of 'taking the bus without wearing any mouth covering' vs. 'wearing a scarf instead of a mask'). As argued in *section 3*, this can be very insightful, as it would inform the researcher about the structure of social incentives that sustain the norm of using masks in public transport (e.g., consider the case where all other available behaviors—wearing no mask, wearing a scarf, etc.— are considered 'very inappropriate' vs. the case where these alternative behaviors are also considered 'mildly inappropriate').

The *belief survey* method is simple and easy to implement, especially in settings where monetary incentives are difficult to administer and where there are constraints on the length of the questions to elicit norms (the other elicitation methods, reviewed below, require considerably more time to be administered). However, the method also presents a number of disadvantages that are related to the *lack of monetary incentives* in eliciting responses. Reporting opinions about one's own personal views is easier than reporting beliefs about others' views, as our own opinions are more readily available to us, but predicting others' may require more cognitive effort. Because cognitive effort is costly to exert, the lack of incentives may entice respondents to simply report their first-order beliefs about the situation, rather than their second-order beliefs. Thus, the method may fail to deliver an elicitation of the social norm, despite the questions being explicitly formulated to ask respondents about their second-order beliefs.[11]

11 Another reason why the lack of incentives may undermine the elicitation procedure is the so-called *false consensus bias*, i.e., the tendency for people to believe that their own behaviors and

Another reason why the lack of incentives may undermine the elicitation of norms using the belief survey method is that, in some cases, respondents may be tempted to respond in a socially desirable way to the questions about the appropriateness of certain behaviors. For instance, respondents may wish to describe as relatively appropriate those behaviors that they themselves often take in their daily lives. To go back to the mask example, imagine someone who never wears masks in public transport. The person may well know that others consider this behavior socially inappropriate. However, when asked in a survey, the person may (mis)report that she believes that others find not wearing a mask appropriate, so as to justify her own behavior. This issue can be particularly severe when the researcher uses the same survey to collect information on respondents' own behavior, in which case the motive for socially desirable responding may become stronger.[12] Incentives can constrain this tendency, as they increase the cost of giving socially desirable responses.[13]

Note that there is a subtle, yet worrying relation between inaccurate and social desirable responses, suggesting that one of the two is always likely to occur. Since survey participants are not given any extrinsic motivation to answer accurately, only their intrinsic motivation to help the researcher may provide an incentive to think hard about their responses. Respondents with *weaker* intrinsic motives may spend less cognitive effort to report their second-order beliefs, thus providing inaccurate responses. Respondents with *stronger* intrinsic motives to help the researcher may, however, be more prone to give socially desirable responses.

Overall, these considerations suggest that the lack of incentives may increase the noisiness of responses collected with the belief survey method. It is plausible that the problem may be more serious in settings where the researcher is investigating norms that concern behaviors that are either particularly difficult to evaluate (e.g., situations in which the norm may not be very clear so that it is cognitively more costly to think about it, as it could well be the case for the mask-wearing exam-

opinions are relatively common among the wider population. The use of incentives may reduce the impact of this bias, as incorrectly overweighing one's own opinion becomes monetarily costly.

12 See Rustichini/Villeval 2014 for an illustration of the power of socially desirable responding in the context of normative decisions.

13 Social desirability bias can also take a different form, whereby respondents try and give responses that they think will make them look good in the eyes of the researcher. For instance, in the mask wearing vignette, if respondents believe that the researcher thinks that one should say that not wearing a mask is very inappropriate (e.g., because the researcher comes from a health organization that promotes the use of masks), they may report so even if in truth the norm is more forgiving towards not using a mask in public transport.

ple above), or particularly sensitive (e.g., sexual behaviors), so that the temptation to give socially desirable responses may be higher.

4.2 The 'Krupka-Weber' Method

Two economists, Erin Krupka and Roberto Weber (2013), have developed a procedure to elicit second-order beliefs about social appropriateness using monetary incentives. As in the *belief survey* method, the researcher presents respondents with a vignette describing various possible behaviors that a person may take and asks their beliefs about the appropriateness of these behaviors. The differentiating feature of the *Krupka-Weber* is that respondents are paid monetary rewards if they rate the appropriateness of behavior*in the same way as most other respondents* do.

In the mask-wearing vignette example described above, respondents rate the appropriateness of the person's behavior in the vignette on a Likert scale knowing that they will receive a monetary reward only if they rate the behavior in the same way as most other people who are also evaluating the same vignette. The incentives transform the game into a pure coordination game where respondents (players) have to tacitly coordinate with others in the way they rate behavior in the vignette. To solve the game, players have to find a way to coordinate, and Krupka and Weber argue that social norms can act as a very salient coordination device in the task: to choose the same rating as others, a player can simply refer to her second-order beliefs about what most others consider appropriate or inappropriate, and rate the actions accordingly. If everyone follows the same strategy, then the *Krupka-Weber* method indirectly reveals what the norm is, as respondents' ratings will reflect their second-order beliefs about what is appropriate behavior in the situation described in the vignette.

The use of incentives constitutes an advantage of the *Krupka-Weber* method over the *belief survey* method, since respondents have now a clear extrinsic motivation to think hard about their second-order beliefs and to resist any responding bias that may tempt them to misreport these beliefs. The evidence discussed in Erkut (2020) indeed shows that the *Krupka-Weber* method does not suffer much from responding biases, especially compared to non-incentivized methods (see also D'Adda et al. 2016).

However, the method has also been criticized for a number of potential disadvantages. Perhaps paradoxically, one important critique revolves exactly around the use of monetary incentives to coordinate with others. While it is *plausible* that respondents use the social norm (if one exists) to coordinate, any other feature of the survey or the vignette may serve as a coordination device. For instance, respondents could follow a coordination strategy that specifies to always pick the

left-most rating on the Likert scale.[14] If certain features are particularly salient, respondents may actually prefer to use these alternative coordination devices rather than the social norm, as the salience may increase their chances to get the monetary reward.

Fallucchi/Nosenzo (2020) study this issue empirically, by purposely introducing salient alternative coordination devices in the *Krupka-Weber* method. To do so, they attach differently-sized visual labels to the Likert scale that respondents use to rate actions (see Figure 1). To test the extent to which the inclusion of the labels distorts appropriateness ratings, they compare responses in the task with visual labels to responses in a standard *Krupka-Weber* task without visual labels. For instance, in the case of the labels shown in Figure 1, one would expect that, if respondents use the labels to coordinate and recognize the oversized green smileys as a salient coordination device, there should be an inflation of 'Very socially appropriate' ratings in the task with visual labels compared to the task without.

Very socially inappropriate	Somewhat socially inappropriate	Somewhat socially appropriate	Very socially appropriate

Tab. 1: Visual labels used in Fallucchi/Nosenzo (2020)

Fallucchi/Nosenzo (2020) find that ratings are substantially robust to the inclusion of visual labels: respondents generally do not rate actions differently when there is a salient alternative coordination device available in the task. There is an exception, though. In situations where there is *not* a clear norm to start with, then there are significant distortions in the ratings of appropriateness: respondents tend to pick

14 As another example, consider a survey where the researcher wants to find out whether people think that expressing approval of xenophobic views is inappropriate. To do this, the researcher constructs a vignette where two people discuss whether making xenophobic remarks on social media is acceptable, and one person says 'I think it is always very appropriate for people to freely express their opinion, even when it is racist'. Imagine the researcher asks respondents to rate the appropriateness of that statement. Respondents may think that the statement is actually inappropriate. However, because the statement contains the words 'very appropriate', they may rate the statement as very appropriate, anticipating that other respondents would do the same in order to secure the monetary reward.

more often the rating with the salient label attached, relative to a task without labels. These results suggest that, although in theory the *Krupka-Weber* method can be vulnerable to the presence of non-norm-related coordination devices, in practice this problem does not seem to have much bite—unless the decision situation the researcher is studying is one where there is not a clear norm that regulates behavior.

Another potential disadvantage of the *Krupka-Weber* method (that also extends to the *belief survey* method) is that the focus on second-order beliefs, while conceptually in line with the notion of norms, does not allow to study phenomena such as *pluralistic ignorance*. As discussed in *section 3.2*, this can lead to situations where individuals perceive norms that no one really endorses privately. Knowing that a situation is characterized by pluralistic ignorance can be useful. Norms that are not endorsed by large sections of a society and that only exists because of miscalibrated beliefs are—in theory—easier to dispel (e.g., via suitable informational campaigns), compared to norms that are supported by a wide consensus in the population.[15]

It is worth discussing two further concerns about the *Krupka-Weber* method. The first concern is conceptual. Although we emphasized the notion of social norms as second-order beliefs about appropriateness, it is not completely clear whether these are in fact the beliefs that the *Krupka-Weber* method elicits. In order to coordinate, a player has to form beliefs about how other players will rate the actions in the vignette. Assume that the player thinks that others will use the social norm to rate the actions, i.e., that other players will use their (second-order) beliefs about what others believe is appropriate. Then the player will rate the actions according to what she believes other players believe others believe is appropriate. This, however, is a *third*-order belief (the player's belief about others' second-order beliefs), and not a second-order belief. Indeed, the logic can be iterated further, e.g., the player may think that others use their third-order beliefs to rate the actions in the task, and so base her choices on her fourth-order belief, and so on. Thus, conceptually, it is not clear what order of beliefs the *Krupka-Weber* method actually elicits.

The second concern is of technical nature. The difficulty of the Krupka-Weber coordination game increases proportionally with the number of possible appro-

15 Galbiati et al. 2020 offer an interesting example of how laws can dispel norms sustained by pluralistic ignorance. Focusing on the UK during the 2020 COVID-19 pandemic, they show that, before the UK government enacted a law putting the country under lockdown, most people personally approved the adoption of behaviors like social distancing or not shaking hands, but at the same time believed that most other people did not approve it. The enactment of the lockdown law, however, quickly dispelled the (miscalibrated) second-order beliefs, putting them in line with people's first-order beliefs (that were unchanged by the law).

priateness ratings that respondents can choose from. A coordination game with only two possible ratings ('appropriate' or 'inappropriate') is much easier than a game with six ratings ('very appropriate', 'somewhat appropriate', 'appropriate', 'inappropriate', 'somewhat inappropriate', 'very inappropriate'). As the game becomes more difficult, respondents' incentive to think hard about others' beliefs may become smaller. To avoid this, applications of the Krupka-Weber method commonly deploy a four-point Likert scale ('very appropriate', 'somewhat appropriate', 'somewhat inappropriate', 'very inappropriate). This, however, constrains the effectiveness of the elicitation procedure in detecting subtle differences between actions.[16] There seems thus to be a trade-off between precision of the measurement and power of the incentives, which, in some cases, may limit the ability of the researcher to identify small but systematic differences in norms across different subgroups or contexts.[17]

4.3 The 'Opinion Matching' Method

A third norm-elicitation method has been used in the literature that addresses many of the concerns raised against the *belief survey* and *Krupka-Weber* methods (e.g., Bicchieri/Xiao 2009; Bicchieri et al. 2019). We call this approach the *opinion matching* method. It consists of a two-step elicitation procedure. In a first step, a group of respondents are asked to report their opinion about the appropriateness of behavior in a vignette. In the second step, a new or the same group of respondents (depending on whether the elicitation is done between- or within-subject), are asked to guess the most common response of the first group (i.e., to report their second-order beliefs of appropriateness), and are given monetary incentives proportional to the accuracy of their guesses.

16 To see this, consider the case of a two-point Likert scale ('appropriate' or 'inappropriate'). Imagine the researcher wants to study whether there are differences in the rating of an action between men and women. Small systematic differences in their appropriateness judgments may exist (e.g., men may think the action is very appropriate, while women may think it is moderately appropriate). However, since participants are constrained to use only two ratings, their observable responses will be the same.

17 Merguei et al. 2020 provide a possible solution to this problem. They introduce a 'continuous' norm elicitation task where subjects are not constrained to a discrete Likert-scale in rating the appropriateness of actions. To incentivize the elicitation, Merguei et al. pay each subject based on the proportion of other participants that submit a rating in a neighborhood of the subject's rating, which mitigates the issue of the difficulty of coordinating in games with a large number of possible appropriateness ratings.

Consider the mask wearing vignette that we used as our running example. The researcher could survey two groups of respondents.[18] Respondents in the first group are asked to read the vignette and to simply report their opinion about the appropriateness of the various possible behaviors that the person in the vignette could take (ride the bus without a mask, wear a scarf instead of the mask, go back home to fetch the mask, etc.). In each case, respondents rate the appropriateness of behavior using a Likert-scale. After the data from the first group has been collected, the researcher surveys a second group. Respondents in the second group also read the vignette and learn about the possible actions that the person in the vignette can take. Moreover, they learn that a first group of respondents have already been interviewed and have provided appropriateness ratings for the behaviors in the vignette. For each behavior, the respondents of the second group are asked to guess the most frequent appropriateness rating provided by the first group. For instance, they would be asked to guess the most common appropriateness rating for the action 'ride the bus without a mask' among the first group of participants and earn a monetary reward that is proportional to how accurate their guess is.

This *opinion matching* method has a number of advantages relative to the two methods discussed above. First, the use of incentives in the elicitation of second-order beliefs (from the second group of respondents) is an improvement over the belief survey method for the reasons mentioned in the previous subsection (extrinsic incentives motivate respondents to think harder about their beliefs and to override any responding biases). Second, the sequential elicitation of first and second-order beliefs removes the strategic component that is present in the coordination game of Krupka and Weber, thus eliminating the potential for measurement distortions due to the use coordination devices that are extraneous to the social norm.[19] Third, contrary to the *Krupka-Weber* method, the *opinion matching* method has a clear focus on second-order beliefs. In fact, when run within-subject, it may even help respondents to better appreciate the difference between first-order beliefs (what they personally consider appropriate, asked in the first step) and second-order beliefs (what most others consider appropriate, asked in the second

18 We here give an example of a between-subject design, but the method could just as easily been run within-subject, by going through the two sequential steps with the same group of respondents. In fact, as discussed below, when conducted within-subject, the method may have some additional advantages over alternative methods.

19 In the opinion matching method respondents are guessing the responses of other people who have *already* completed the task without knowing about the two-step elicitation procedure (i.e., respondents in the first group do not know that there will be a group of respondents asked to guess their opinions). This removes any incentive to respond strategically to coordinate with others. The only objective for subjects in the second group is to accurately guess the first group's responses.

step), and thus could increase the general quality of responses.[20] Finally, another advantage of the two-step method is that the researcher obtains data on both first- and second-order beliefs and can thus check, for instance, whether there are significant discrepancies between the two beliefs, as an indication of pluralistic ignorance.[21]

However, the *opinion matching* method has also some disadvantages. The most serious concern is perhaps the fact that the first-step responses are not incentivized (unavoidably, since they measure personal opinions). This means that these responses could be vulnerable to responding biases, potentially even more so than the responses collected with the *belief survey* method since, in this case, they concern the respondent's *own* beliefs about the appropriateness of specific behaviors. Note that this could undermine not only the measurement of first-order beliefs, but also that of second-order beliefs, since respondents are in fact incentivized to guess others' first-order beliefs. If these first-order beliefs are distorted (e.g., due to socially desirable responding), and if second-step respondents anticipate this, then their guesses (second-order beliefs) will reflect the belief distortion that occurred in the first step. This issue may become particularly severe in case of within-subject designs, where, in the second step, respondents may be more easily aware of any distortions in the responses they provided in the first step.[22]

Another potential disadvantage of the opinion matching method is that it is slightly more cumbersome to deploy in empirical studies (especially compared to the belief survey method), because it requires assigning response-contingent payments to respondents and either recruiting two separate groups of subjects (between-subject designs) or conducting two iterations of questions with the same

20 Respondents may find it easier to understand the notion of second-order beliefs once they have been asked about their first-order beliefs. In practice, in the second step the researcher has to remind respondents about the appropriateness questions they were asked in the first step, and then ask them to guess what most other survey participants have responded to that same question.

21 More generally, some research questions may require the elicitation of both first- and second-order beliefs, e.g., if one is interested in studying the effects of a policy intervention on both norms and personal values.

22 Consider, for example, a researcher interested in studying norms about sexual behavior. Imagine the researcher uses a within-subject design. In the first step, respondents may report opinions that are socially desirable even if they do not reflect their actual opinion (e.g., they may report that the use of contraceptives for premarital sex is very appropriate, although they actually think it is not and know that most other people also think it is not). In the second step, when asked to guess what most other people have answered in the first step, respondents may anticipate that others, like themselves, may have reported to the researcher a distorted version of their first-order beliefs. Therefore, they may report that they think that most others believe that the use of contraceptives is appropriate, while knowing that, in truth, no one believes this and that there is no norm favouring the use of contraceptives in their society.

group of respondents (within-subject designs). This may also increase the monetary costs of the survey.

4.4 Call for Further Research

Before turning to some general conclusions, it is useful to discuss a few remaining issues that are common to the three methods and highlight the need for further research in some specific areas. First, it is important to note that the methods discussed here are concerned with the measurement of (injunctive) norms—but they do not say anything about the extent to which individuals are actually *willing to follow* these norms (on this point, see also the excellent discussion in Vostroknutov 2020).

To address this point, many of the studies that elicit norms, often also run complementary experiments or surveys to collect data on actual behavior—so that they can compare whether the actions that people take in a given situation are consistent with the norms that prevail in that situation. However, the behavioral mechanisms that make individuals more or less prone to follow norms are not yet fully understood (e.g., Are there systematic factors that make some individuals more or less prone to norm-following? Does the preference for following norms depend on observation of norm-following or norm violations by others? Etc.).

Some recent advances in this area have been made by Kimbrough/Vostroknutov (2016) and Gächter et al. (2020). These studies introduce tasks that allow to measure the individual preferences to comply with norms and to observe whether these preferences are conditional on what others do and believe ought to be done (which some models explicitly assume, e.g., Bicchieri 2006, while others explicitly rule out, e.g., Kimbrough/Vostroknutov 2016). We think that further research in this direction, combining and tying together measurements of (i) preferences for norm-following, (ii) injunctive norms, and (iii) actual behavior, is very promising, as it will allow to shed light on the functioning of models of norm-compliance and, ultimately, on their ability to explain human behavior.

A second area where further research is needed concerns the interpretation of norms data in settings where the distribution of second-order beliefs does not clearly converge towards a consensus for what is appropriate or inappropriate behavior. This can happen for a number of reasons. In some cases, respondents' ratings may show a large degree of dispersion. For instance, this is the case in standard dictator games for hypergenerous actions that leave the recipient with more money than the dictator, as mentioned earlier. Typically, there is a large heterogeneity in how these actions are rated, with half of respondents considering them appropriate, and the other half inappropriate. A lack of convergence towards

a clear norm may also occur if all actions available to a decision-maker are rated in a similar way, so that, for instance, all actions are rated as appropriate. In either case, the interpretation of data does not seem straightforward. On the one hand, one may view these data patterns as indicating that a social norm does *not* exist in these situations, since, by definition, a social norm is a *shared* understanding about the appropriateness of actions. However, alternative interpretations seem possible. In the case of heterogeneous rating, for example, a possible interpretation is that there may be a *multiplicity of normative principles* being recognized by different subgroups of the population.

Fromell et al. (2019) provide an example. An argument that has been proposed in the literature to explain why individuals from poor communities in Sub-Saharan Africa fail to accumulate wealth is that in these communities there are strong sharing norms that force individuals to share with others any wealth that they manage to accumulate. Fromell et al. (2019) conduct a series of lab-in-the-field experiments in rural Kenya to find tangible evidence of these *sharing norms*. On the aggregate, they indeed find that individuals from poor communities consider it appropriate to share most of one's wealth with other village members. However, their data also show considerable dispersion in the normative ratings, with a sizeable fraction of respondents actually finding sharing with others *inappropriate* instead of appropriate. To make sense of this heterogeneity, Fromell et al. conduct a hierarchical cluster analysis that uses an algorithm to organize the data patterns in separate groups ('clusters') that are internally homogeneous and distinct from one another. The analysis reveals that the aggregate heterogeneity occurs because different respondents seem to follow different normative principles, some supporting a strong sharing norm, but others actually supporting the reverse norm, whereby accumulating wealth is viewed as appropriate. Interestingly, Fromell et al. also find that these different norms are supported by systematically different subgroups of the population. The traditional sharing norm, for instance, is supported by individuals who are significantly older and poorer. These systematic group differences lend credibility to the interpretation that the heterogeneity in aggregate data actually masks a plurality of normative principles, supported by different cliques of individuals. Thus, although one may conclude that a *global* norm does indeed not exist in this setting, there is also evidence that different *local* norms prevail within different strata of the population.

This example illustrates how focusing just on aggregate appropriateness ratings may preclude interesting observations about the underlying norms. There is however a difficulty in moving beyond aggregate data: when can we say that a social norm exists or does not exist, based on the distribution of second-order beliefs that we obtain from either of the elicitation methods discussed above? We see value

in pursuing further research in this direction, both at a conceptual/theoretical level and an empirical level.

5 Recommendations and Conclusions

In this paper, we argued that understanding social norms is important for developing more accurate theories to explain human behavior and, ultimately, for designing effective policies that can spur behavioral change. The examples we discussed highlight why, in order to gauge the impact of norms on human decision-making, it cannot be enough to study behavioral regularities alone. Instead, we need to obtain direct measures of *injunctive norms* (what one ought to do), through *second-order beliefs*, and over a *range of behaviors*. To offer some practical guidance, we have discussed the state of the art in economic research, which features three elicitation methods to obtain measures of norms in consideration of these three criteria. We conclude the paper with a list of hands-on recommendations for researchers who are interested in measuring social norms and have to choose one of the three methods.

As regards the choice of a specific method to measure norms, we think that, whenever it is possible to administer incentives to respondents, the *Krupka-Weber* method or the *opinion matching* method should be preferred to the (non-incentivized) *belief survey* method. The theoretical drawbacks of the lack of incentivization discussed earlier are sufficiently serious to potentially compromise the accurate elicitation of norms data, and hence one of the two methods with incentives may be preferable.[23]

A second point to consider is how sure the researcher is that a clear social norm exists. In settings where she is not sure, or where the research design may imply that alternative coordination points may exist in the *Krupka-Weber* task, it

23 However, the extent of these distortions is an empirical matter, for which there is not a lot evidence. Veselý 2015 is the only study we are aware of and that compares incentivized and non-incentivized measurements of norms (using the Krupka-Weber method). He finds only small differences in elicited norms in the ultimatum game. Rustichini/Villeval 2014, on the other hand, find evidence that respondents distort their (unincentivized) moral judgments to justify their behavior. Another potential caveat against the use of incentives in eliciting norms is that incentives may induce respondents to think in a different way (e.g., more materialistically) about morals and norms, compared to situations in which there are no incentives. While we are not aware of any direct evidence of this (and the Veselý paper cited above may be even taken as evidence to the contrary), there is more general experimental evidence that focusing participants on money may alter their motivation and behavior (e.g., Vohs et al. 2006).

may be preferable to use the *opinion matching* method. As an example, consider Lane et al. (2020). They study the influence of laws on social norms by eliciting normative judgments about behaviors that fall on either side of a legal threshold. For instance, to assess the impact of a speed-limit law on the norm of driving, Lane et al. ask respondents to rate the social appropriateness of driving at a speed that is either just below or just above the legal speed limit. One concern with the *Krupka-Weber* task in this setting is that, instead of using the social norm, respondents may use the legal threshold as a way to coordinate in the game. For example, they could use the following strategy: to rate all actions that fall on the legal side of the threshold as appropriate and all actions that fall on the illegal side as inappropriate, *regardless* of what the social norm actually says about these actions. Although Lane et al. find little evidence of distortions due to this alternative coordination strategy, an elicitation procedure that does not rely on coordination games to measure norms may be preferable in such settings, as it completely circumvents the issue of strategic coordination.

Finally, the *opinion matching* method—with its combined elicitation of first- and second-order beliefs—offers a further advantage over the other methods, in that it delivers data that allow to study whether there is a misalignment between respondents' personal normative beliefs and the social norm, as in the case of pluralistic ignorance. Note, however, that while this is a feature that is automatically built in the opinion matching method, the other two methods could be easily extended to allow for the measurement of first-order beliefs in conjunction with second-order beliefs.[24]

References

Abeler, J./D. Nosenzo/C. Raymond (2019), Preferences for Truth-Telling, in: *Econometrica* 87(4), 1115–1153

d'Adda, G./M. Drouvelis/D. Nosenzo (2016) Norm Elicitation in Within-subject Designs: Testing for Order Effects in: *Journal of Behavioral and Experimental Economics* 62, 1–7

— /M. Dufwenberg/F. Passarelli/G. Tabellini (2020), Social Norms with Private Values: Theory and Experiments, in: *Games and Economic Behavior* 124, 288–304

Akerlof, G.A./R.E. Kranton (2000), Economics and Identity, in: *The Quarterly Journal of Economics* 115(3), 715–753

24 Barr et al. (2020), for instance, use two versions of the Krupka-Weber method, one with incentives to coordinate as in the original method, and one without incentives where respondents are simply asked to report their individual normative beliefs (i.e., their first-order beliefs of appropriateness).

Andreoni, J./B.D. Bernheim (2009), Social Image and the 50–50 Norm: A Theoretical and Experimental Analysis of Audience Effects, in: *Econometrica* 77(5), 1607–1636

Barr, A./M. Dekker/F. Mwansa/T.L. Zuze (2020), *Financial Decision-Making, Gender and Social Norms in Zambia: Preliminary Report on the Quantitative Data Generation, Analysis and Results*. CeDEx Discussion Paper 2020-06

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

Bašić, Z./Verrina (2020), *Personal Norms, Social Norms and Image Concerns*. Working Paper under preparation

Beblo, M./L. Görges (2018), On the Nature of Nurture. The Malleability of Gender Differences in Work Preferences, in: *Journal of Economic Behavior & Organization* 151, 19–41

Bénabou, R./J. Tirole (2006), Incentives and Prosocial Behavior, in: *American Economic Review* 96(5), 1652–1678

Bernheim, B.D (1994), A Theory of Conformity, in: *Journal of Political Economy* 102(5), 841–877

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

— (2017), *Norms in the Wild*, New York

— /E. Dimant/S. Gaechter/D. Nosenzo (2019), *Observability, Social Proximity, and the Erosion of Norm Compliance*. SSRN Discussion Paper 3355028. Rochester/NY

— /E. Xiao (2009), Do the Right Thing: But Only if Others Do So, in: *Journal of Behavioral Decision Making* 22(2), 191–208

Blau, F.D./L.M. Kahn (2013), Female Labor Supply: Why Is the United States Falling Behind? In: *American Economic Review* 103(3), 251–256

Bursztyn, L./A.L. González/D. Yanagizawa-Drott (2020), Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia, in: *American Economic Review* 110(10), 2997–3029

Erkut, H. (2020), Incentivized Measurement of Social Norms Using Coordination Games, in: *Analyse & Kritik* 42(1), 97–106

European Commission (2017), Special Eurobarometer 465: Gender Equality. Technical Report and data available online

— (2020), A Union of Equality: Gender Equality Strategy 2020-2025. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, available online

European Institute for Gender Equality (2019), Gender Equality Index 2019. Work-Life Balance. European Institute for Gender Equality publication, available online

Fallucchi, F./D. Nosenzo (2020) *The Coordinating Power of Social Norm*s. CeDEx Discussion Paper 2020-14

Fehr, E./I. Schurtenberger (2018), Normative Foundations of Human Cooperation, in: *Nature Human Behaviour* 2(7), 458–468

Forsythe, R./J.L. Horowitz/N.E. Savin/M. Sefton (1994), Fairness in Simple Bargaining Experiments, in: *Games and Economic Behavior* 6(3), 347–69

Fromell, H./D. Nosenzo/T. Owens/F. Tufano (2019), *One Size Doesn't Fit All: Plurality of Social Norms and Saving Behavior in Kenya*. CeDEx Discussion Paper 2019-12

Gächter, S./L. Gerhards/D. Nosenzo (2017), The Importance of Peers for Compliance With Norms of Fair Sharing, in: *European Economic Review* 97(C), 72–86

— /L. Molleman/D. Nosenzo (2020), *Why Do People Follow Rules?* Working Paper under preparation.

Galbiati, R./E. Henry/N. Jacquemet/M. Lobeck (2020), *How Laws Affect the Perception of Norms: Empirical Evidence from the Lockdown*. CEPR Discussion Paper No. DP15119, Available at SSRN: https://ssrn.com/abstract=3674895. Rochester/NY

Givati, Y./U. Troiano (2012), Law, Economics, and Culture: Theory of Mandated Benefits and Evidence from Maternity Leave Policies, in: *The Journal of Law & Economics* 55(2), 339–364

Gneezy, U./S. Saccardo/R. van Veldhuizen (2019), Bribery: Behavioral Drivers of Distorted Decisions, in: *Journal of the European Economic Association* 17(3), 917–946.

Goldin, C. (2006), The Quiet Revolution That Transformed Women's Employment, Education, and Family, in: *American Economic Review* 96(2), 1–21

Görges, L. (2019), *Wage Earners, Homemakers & Gender Identity – Using an Artefactual Field Experiment to Understand Couples' Labour Division Choices*, Working Paper

— /D. Nosenzo. (2020), Social Norms and the Labor Market, in: *Handbook of Labor, Human Resources and Population Economics*, ed by. Klaus F. Zimmermann, 1–26, Cham

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

Kliemt, H. (2020), Economic and Sociological Accounts of Social Norms, in: *Analyse & Kritik* 42(1), 41–96

Krupka, E./R.A. Weber (2013), Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary? In: *Journal of the European Economic Association* 11(3), 495–524

Lane, T./D. Nosenzo/S. Sonderegger (2020), *Law and Norms: Empirical Evidence*. Working paper under preparation

Merguei, N./M. Strobel/A. Vostroknutov (2020), *Moral Opportunism and Excess in Punishment Decisions*. Working Paper under preparation

OECD (2017), The Pursuit of Gender Equality: An Uphill Battle, available online

Reuben, E./A. Riedl (2013), Enforcement of Contribution Norms in Public Good Games with Heterogeneous Populations, in: *Games and Economic Behavior* 77(1), 122–137

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

Sugden, R. (1998), Normative Expectations: The Simultaneous Evolution of Institutions and Norms, in: *Economics, Values, and Organization*, ed by. A.Ben-Ner/L.G. Putterman, Cambridge

– (2000), The Motivating Power of Expectations, in: *Rationality, Rules, and Structure*, ed by. J. Nida-Rümelin/W. Spohn, 103–129, Dordrecht

Veselý, Š. (2015), Elicitation of Normative and Fairness Judgments: Do Incentives Matter? In: *Judgment and Decision Making* 10(2), 191–197

Vohs, K.D./N.L. Mead/M.R. Goode (2006), The Psychological Consequences of Money, in: *Science* 314(5802), 1154–1156

Vostroknutov, A. (2020), Social Norms in Experimental Economics: Towards a Unified Theory of Normative Decision Making, in: *Analyse & Kritik* 42(1), 3–40