Judgment and Decision Making, Vol. 10, No. 2, March 2015, pp. 191–197

Elicitation of normative and fairness judgments: Do incentives matter?

Štěpán Veselý*†

Abstract

Krupka and Weber (2013) introduce an incentive-compatible coordination game as an alternative method for elicitation of normative judgments. I show, however, that people provide virtually the same responses in incentivized and non-incentivized versions of the Krupka-Weber game. Besides ratings of social appropriateness, I also elicit ratings of fairness of all possible offers in an ultimatum game. Ratings of social appropriateness and fairness are similar for low offers (below or equal to the equal split), but not for high offers which are judged to be more appropriate than fair.

Keywords: appropriateness, fairness, social norms, incentives, rating, ultimatum game.

1 Introduction

The study of social norms is becoming common in behavioral economics (Bardsley & Sausgruber, 2005; Croson, Handy & Shang 2009; Bicchieri & Xiao, 2009; Krupka & Weber, 2009, 2013; Reuben & Riedl, 2013). This can be seen as a continuation and extension of behavioral economists' interest in social preferences, such as fairness or altruism (e.g., Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000; Andreoni & Miller, 2002; Cappelen et al., 2007).

While both preferences and norms can be, to an extent, inferred from observed behavior, eliciting perceived social preferences and norms directly from subjects can be another useful way of collecting data. However, economists often view self-report measures as problematic, because they are in theory just cheap talk: the experimenter (often) cannot rule out the possibility that respondents are not motivated to provide truthful and accurate responses unless he himself provides them with incentives to do so (Camerer & Hogarth, 1999; Hertwig & Ortmann, 2001).

Making subjects' responses incentive-compatible can be especially challenging when eliciting normative or fairness judgments, as "objective" criteria against which to compare the responses are not obvious. To deal with this methodological obstacle, Krupka and Weber (2013) introduce the first incentive-compatible method for elicitation of social norms (the KW game henceforth). A test of this method will be the focus of this paper.

The KW game is a pure matching coordination game in which each participant receives a monetary reward if his or her rating of social appropriateness of a randomly chosen action matches the modal rating of this action's appropriateness in his or her experimental session. Shared social norms are assumed to create focal points for the judgment of appropriateness on which participants can coordinate (Krupka & Weber, 2013).

The KW game has been used to measure perceived social appropriateness of actions in various games, such as the dictator game and gift-exchange game (Gächter et al., 2013; Krupka & Weber, 2013; Erkut et al., 2014; Krupka et al., 2014). Burks and Krupka (2012) use the KW game to identify norms for on-the-job behavior among employees at a financial services company.

However, the aforementioned studies do not attempt to show that the KW game actually leads to different ratings of social appropriateness than a comparable incentive-free rating method. In this paper, I therefore provide a comparison of the KW game to such a non-incentivized, but otherwise very similar rating method.[1]

As mentioned above, in the experimental economics tradition material incentives are commonly used to mo-

I would like to thank the editor, Jon Baron, and two anonymous referees for valuable comments and suggestions. I gratefully acknowledge financial support from Masaryk University through grant MUNI/M/0045/2013 "Experimental analysis of repeated choice: Economic and political science approaches".

*Department of Psychology, Faculty of Arts, Masaryk University, Arne Novaka 1, Brno, 602 00, Czech Republic. E-mail: 146287@mail.muni.cz.

†Department of Public Economics, Faculty of Economics and Administration, Masaryk University

[1] Xiao & Houser (2005) and Houser & Xiao (2011) use a game very similar to the KW game for classification purposes. Houser & Xiao (2011) find that coders who are financially rewarded for successful coordination on shared meanings of messages they classify are more responsive to classification criteria than coders in a traditional unincentivized content analysis treatment. However, as the authors mention: "Content analysis differs from our coordination game in multiple ways (e.g., evaluating according to what one thinks others believe as compared to what oneself believes)" (p. 7). Thus, the presence of incentives is only one possible explanation of the observed differences between coders' performance across the two treatments in Houser and Xiao's experiment.

https://doi.org/10.1017/S1930297500003958 Published online by Cambridge University Press

tivate decisions. However, in some cases cash incentives do not seem to make a difference. For example, meta-analyses by Zelmer (2003) and Engel (2011) show that the presence (vs. absence) of pecuniary incentives in public goods games and dictator games, respectively, does not significantly predict decisions. In their survey article, Camerer and Hogarth (1999) show that the effects of incentives tend to be "mixed and complicated" (p. 8) and "depend dramatically on the kind of task a person performs" (p. 20). It is therefore reasonable to ask whether incentivizing "correct" (i.e., modal) responses in the KW game has an effect.

So far, the KW game has been used to rate social appropriateness of actions (see citations above). However, its use can be easily extended to elicit shared judgments of fairness and other aspects of behavior. Hence, participants in my study rate social appropriateness and fairness of possible actions in the ultimatum game (UG, Güth et al., 1982). It is quite possible that incentivization will matter when eliciting one type of judgment, but not the other, for example because one type of judgment could be more difficult to make.

Fairness is often studied using UG and its variants. Still, experimental participants are seldom asked about their actual fairness perceptions, as noted by van Winden (2007). The KW game enables elicitation of impartial fairness perceptions of different actions in UG in an incentive-compatible way. This has probably never been done before: for example in Leliveld et al. (2009), Reuben and van Winden (2010) and Rustichini and Villeval (2014) participants rated fairness of actions in UG or similar games, but these ratings were not incentive-compatible.

2 Method

270 university students from all fields of study (173 females) participated in the study across 16 sessions, each of which lasted for approximately 1 h. Participants earned 166.5 CZK (approximately $7.7) on average including the show-up fee.[2] The show-up fee was 90 CZK in the KW game treatment and 120 CZK in the Non-incentivized treatment.

The experiment was implemented as a 2 (elicitation method: KW game vs. Non-incentivized) * 2 (order of presentation: fairness rated first vs. social appropriateness rated first) * 11 (which offer is being rated: $0, $1, ..., $10) mixed design with two dependent variables (fairness rating and appropriateness rating). "Elicitation method" and "order of presentation" are independent factors. Combination of these factors yields four independent conditions to which participants were randomly assigned. "Offer being rated" is a repeated measures factor.

I adapted the instructions from Krupka and Weber (2013) and I ran the experiment in a paper-and-pencil form, as they did[3] (the instructions are in the Supplement).

Participants in all conditions rated social appropriateness and fairness of all possible actions a proposer can take in UG. There were 11 possible actions for the proposer (to offer $0–$10 in $1 increments), hence there were 11 ratings of social appropriateness and 11 ratings of fairness per participant. Fairness and social appropriateness were rated on a four-point scale as e.g., "very unfair", "somewhat unfair", "somewhat fair" and "very fair".

In both the KW game treatment and in the Non-incentivized treatment participants were asked to rate the offers from the position of a third, neutral party that is not involved in the UG. Also, in both treatments, participants were asked to provide what they thought would be a "common" or "most frequent" rating, rather than their "personal" rating.

The key experimental manipulation involved incentives. In the KW game treatment, but not in the Non-incentivized treatment, two possible actions in the UG were randomly (with replacement) selected at the conclusion of the experiment. Participant received a monetary reward (75 CZK) if his/her rating of social appropriateness matched the modal rating of social appropriateness in his/her session for the first selected action, and he/she received (another) 75 CZK if his/her rating of fairness matched the modal rating of fairness in his/her session for the second selected action.

In the Non-incentivized treatment participants rated the choices without incentives. However, they were also asked not to provide their "personal" ratings, but rather "what they think would be 'common' ratings", such as ratings "they think would be the most common among people in today's session". This request in the instructions points to a possible focal point (the modal rating) and thus makes the Non-incentivized treatment as similar to the KW game treatment as possible, except for the incentivization.

3 Results and discussion

Table 1 presents means and standard errors of incentivized and non-incentivized ratings of social appropriateness and fairness of proposer's possible choices. The ratings are coded so that the lowest rating (e.g., "very unfair") has the value equal to 1, while the highest rating has the value equal to 4.

Table 1 shows, for example, that the mean rating of social appropriateness of the lowest possible offer ($0)

[2]The average hourly wage for a student is about 80 CZK ($3.7) in Czech Republic.

[3]Two guessing games were played after the main part of the experiment and a brief cognitive ability test (CRT) was administered. The results from these tasks are not reported here.

https://doi.org/10.1017/S1930297500003958 Published online by Cambridge University Press

Table 1: Incentivized and non-incentivized ratings of proposer's choices: mean (S.E.).

Proposer's offer being rated	Ratings of social appropriateness in the KW game treatment	Ratings of social appropriateness in the Non-incentivized treatment	Ratings of fairness in the KW game treatment	Ratings of fairness in the Non-incentivized treatment
\$0	1.09 (0.03)	1.09 (0.03)	1.07 (0.03)	1.05 (0.03)
\$1	1.42 (0.05)	1.29 (0.05)	1.23 (0.04)	1.22 (0.04)
\$2	1.78 (0.04)	1.57 (0.06)	1.61 (0.05)	1.55 (0.05)
\$3	2.32 (0.06)	2.16 (0.06)	2.07 (0.05)	2.01 (0.06)
\$4	2.98 (0.05)	2.94 (0.05)	2.87 (0.05)	2.79 (0.05)
\$5	3.91 (0.03)	3.91 (0.02)	3.96 (0.02)	3.97 (0.02)
\$6	3.44 (0.06)	3.44 (0.05)	3.09 (0.06)	3.06 (0.06)
\$7	3.14 (0.07)	3.20 (0.07)	2.45 (0.07)	2.30 (0.07)
\$8	2.76 (0.08)	2.83 (0.08)	1.96 (0.07)	1.93 (0.07)
\$9	2.56 (0.09)	2.58 (0.10)	1.73 (0.08)	1.65 (0.08)
\$10	2.48 (0.10)	2.48 (0.11)	1.59 (0.09)	1.50 (0.08)

Note: There were 137 raters in the KW game treatment and 133 in the Non-incentivized treatment.

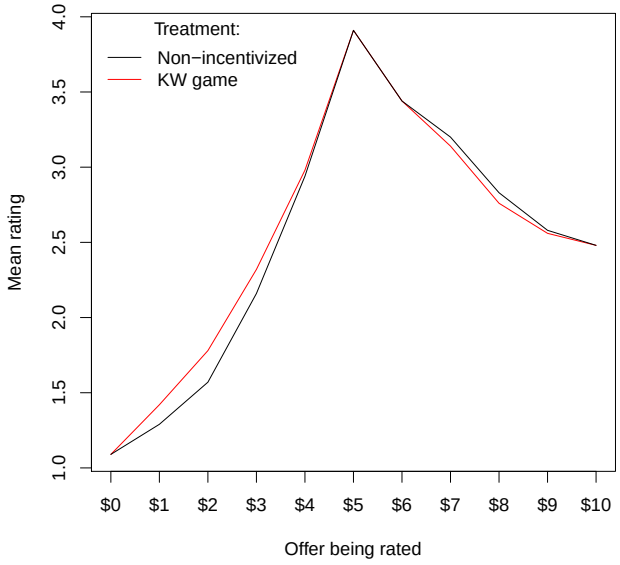
is 1.09 both in the KW game treatment and in the Non-incentivized treatment and that the mean rating of fairness of the \$0 offer is 1.07 in the incentivized KW game treatment and 1.05 in the Non-incentivized treatment. The key comparisons in this paper (to be reported below) will test for differences between ratings summarized in the second vs. the third column in Table 1 (i.e., between appropriateness ratings with vs. without incentives) and between ratings summarized in the fourth vs. the fifth column (i.e., between fairness ratings with vs. without incentives).

The mean social appropriateness and fairness ratings (with and without incentives for raters) are also displayed in Figures 1 and 2, respectively. No important differences between ratings with vs. without incentives are apparent from Table 1 or Figures 1 and 2.

Rather than applying statistical tests to mean ratings of specific offers (e.g., \$0), I first fit four straight lines to each subject's data (as described in the following) and then test for differences between the slopes of the fitted lines. More specifically, this way I obtain four new variables, SlopeApp1 and SlopeApp2 (slopes of lines fitted to ratings of social appropriateness of offers \$0-\$5 and \$5-\$10, respectively) and SlopeFair1 and SlopeFair2 (slopes of lines fitted to ratings of fairness of offers \$0–\$5 and \$5–\$10, respectively). These new variables will be used in subsequent analyses.[4]

Because part of subjects rated social appropriateness first and fairness second and part of subjects the other way round I first tested for the presence of order effects with a series of four t-tests. An independent samples t-test yielded no significant differences in SlopeApp1 between subjects who rated appropriateness first and those who rated it second, $t(268) = 0.240$, $p = .81$, $r = .01$ (negligible effect); mean difference = 0.00 (95% $CI = -0.02$, 0.03). Similarly, I find no significant differences in SlopeApp2 between these two groups of subjects, $t(268) =$

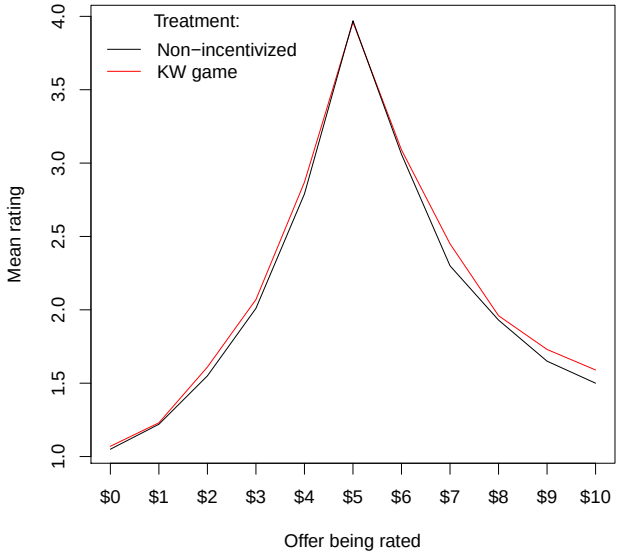
Figure 1: Mean social appropriateness ratings in treatments with and without incentives.



[4] I thank Jon Baron for suggesting this approach to analysis and for helping me with it.

https://doi.org/10.1017/S1930297500003958 Published online by Cambridge University Press

Figure 2: Mean fairness ratings in treatments with and without incentives.



-1.219, $p = .22$, $r = .07$ (very small effect); mean difference = -0.04 (95% $CI = -0.10, 0.02$). Likewise, I find no significant differences in SlopeFair1 between these two groups of subjects, $t(268) = -0.612$, $p = .54$, $r = .04$ (negligible effect); mean difference = -0.01 (95% $CI = -0.03, 0.02$). Finally, I do not find any significant differences between these two groups of subjects in SlopeFair2, either, $t(268) = -0.105$, $p = .92$, $r < .01$ (negligible effect); mean difference = 0.00 (95% $CI = -0.05, 0.05$).[5] Consequently, I pool the data from participants who rated social appropriateness as first and those who rated it as second in all subsequent analyses.

Now I proceed to the main comparisons, which will show whether there is an effect of providing (vs. not providing) incentives to raters on the ratings of social appropriateness and/or fairness.

Using an independent samples t-test I find no significant differences in SlopeApp1 between subjects in incentivized KW-game treatment and subjects in Non-incentivized treatment, $t(268) = -0.681$, $p = .50$, $r = .04$ (negligible effect); mean difference = -0.01 (95% $CI = -0.04, 0.02$).

Similarly, using an independent samples t-test I find no significant differences in SlopeApp2 between subjects in incentivized KW-game treatment and subjects in Non-incentivized control treatment, $t(268) = -0.077$, $p = .94$, $r < .01$ (negligible effect); mean difference = 0.00 (95% $CI = -0.07, 0.06$).

Next, with an independent samples t-test I find no significant differences in SlopeFair1 between subjects in KW-game treatment and subjects in control, $t(268) = 0.152$, $p = .88$, $r < .01$ (negligible effect); mean difference = 0.00 (95% $CI = -0.02, 0.03$).

Finally, with an independent samples t-test I find no significant differences in SlopeFair2 between subjects in KW-game treatment and subjects in control, $t(268) = 0.443$, $p = .66$, $r = .03$ (negligible effect); mean difference = 0.01 (95% $CI = -0.04, 0.06$).

To complement the analyses reported above, I also run two separate three-way mixed ANOVAs, one with social appropriateness rating as dependent variable, the other with fairness rating as dependent variable (the predictors are: the presence of incentives, whether appropriateness or fairness was rated first, and the offer being rated).

The first ANOVA reveals that appropriateness ratings differ across offers, $F(2.38, 630.83) = 483.61$, $p < .001$, effect size $\eta^2 = .65$ (large effect).[6] Bonferroni-corrected pairwise comparisons between offers reveal, for example, that the equal split ($5 offer) is more appropriate than any other offer. Offering $6 is more appropriate than any other offer except $5. Offering $0-$2 is less appropriate than offering any higher offer.[7] Importantly, I found no significant effect of the presence of incentives on social appropriateness ratings $F(1, 266) = 0.75$, $p = .39$, effect size $\eta^2 = .003$ (negligible effect). There was also no significant effect of the order of presentation, $F(1, 266) = 0.50$, $p = .48$, $\eta^2 = .002$ (negligible effect).

The second ANOVA reveals that fairness ratings differ across offers, $F(3.28, 872.67) = 620.08$, $p < .001$, $\eta^2 = .70$ (large effect). Bonferroni-corrected pairwise comparisons between different offers reveal that the equal split is fairer than any other offer. Offering $6 is fairer than any other offer except $5, offering $4 is fairer than any other offer except $5 and $6. Offering $0 or $1 is less appropriate than offering any higher offer.[8] Importantly, I found no significant effect of the presence of incentives on fairness ratings $F(1, 266) = 1.29$, $p = .26$, effect size $\eta^2 = .005$ (negligible effect). There was no significant effect of the order of presentation, either, $F(1, 266) = 0.50$, $p = .48$, $\eta^2 = .002$ (negligible effect).

To conclude, ratings differ across offers, while—which is the main result—incentives do not matter when rating social appropriateness or fairness of actions, at least not in simple bargaining situations such as UG. The effect of the presence of incentives on ratings of both appropriateness and fairness is negligible ($r < .05$ in case of the four t-tests of the incentive effect reported above; $\eta^2 = .002$ for the effect of incentives in both ANOVAs reported above).

[5] These are results of t-tests applied to the entire data set. When performing the same analyses separately on data from treatments with and without incentives, the obtained results are qualitatively similar.

[6] Greenhouse-Geisser adjusted F is used in both ANOVAs because of violation of sphericity.

[7] All reported differences are significant at $p < .05$ (after correction).

[8] All reported differences are significant at $p < .05$ (after correction). Complete results of the multiple comparisons are available upon request.

https://doi.org/10.1017/S1930297500003958 Published online by Cambridge University Press

Based on these results I conclude (as I did earlier) that using incentives virtually does not change elicited normative/fairness judgments. Admittedly, the small effect sizes also mean that I do not have sufficient power to reject the null hypothesis (of no effect of incentives) if it is in fact false. However, I base my conclusion about the minimal effect of incentives not primarily on the significance tests, but rather on the estimated effect sizes.

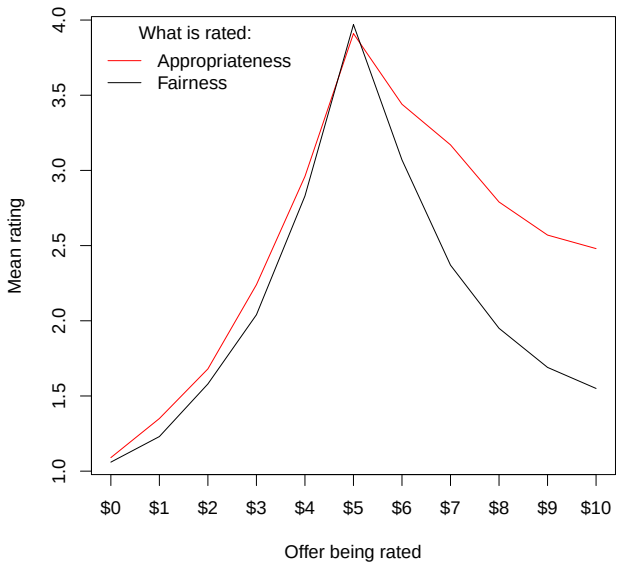
One can think of at least two situations where cash incentives would not affect responses (Camerer & Hogarth, 1999 lay out similar arguments). Consider the following example first. A respondent understands, even absent incentivization, that an action is socially inappropriate. He is not sure, however, whether it is "very" or "somewhat" inappropriate. Offering him incentives for judging the action correctly will not make it any easier for him to discern between the "very" and "somewhat" inappropriate character of that action. This is a situation where incentives would not help because the task is too difficult. Conversely, the effect of incentives could be negligible in very easy tasks where participants do not need any additional motivation to provide the correct response. Thus, it remains an open question whether incentivization of socially shared judgments may not have an effect in more complex settings.

My results suggest that researchers using the KW elicitation protocol might want to follow the "do it both ways" recommendation by Hertwig and Ortmann (2001) and include an unincentivized KW protocol as a control in their future experiments. Consequently, when further evidence accumulates, a more general conclusion concerning the importance of material incentives for normative and fairness judgments can be drawn.

Incentivization is more likely to have an effect on "non-volunteer" participants with low intrinsic motivation to perform well or to provide truthful responses due to, e.g., self-presentation bias. It could be also relatively more effective in supporting performance in time-consuming, lengthy tasks. Finally, could the fact that the raters evaluate a strategic situation (in which they are not involved, however) affect the results? The strategic nature of UG gives the proposer relatively little scope for ignoring distributive norms (see, e.g., Rustichini & Villeval, 2014). It is therefore possible that incentives would be more effective when rating actions in a non-strategic game, such as the dictator game, because distributive norms *might* be generally less clear in the dictator game and stating them might be therefore more costly in terms of cognitive effort on the rater's part.

In the remainder of the paper I at least briefly consider the similarities and differences between appropriateness and fairness ratings. These are apparent in Figure 3. The name of the *y* axis in Figure 3 simply states "Mean rating" which means appropriateness rating in case of the red line and fairness rating in case of the black line. Note that in the left part of Figure 3 the red and black lines are near each other, while in the right part of the figure the red line is above the black one. This indicates that appropriateness ratings are higher than fairness ratings for high offers and about the same for low offers (see the statistical tests below[9]).

Figure 3: Mean social appropriateness and fairness ratings.



A paired *t*-test comparing SlopeApp1 and SlopeFair1 shows only marginally significant, small differences, $t(269) = -1.762$, $p = .08$, $r = .11$ (small effect); mean difference = −0.01 (95% *CI* = −0.02, 0.00). This *t*-test reveals that low offers ($0-$5) receive approximately the same social appropriateness and fairness ratings.

In contrast, there are significant differences between the two types of ratings when high offers are being rated. A paired *t*-test comparing SlopeApp2 and SlopeFair2 shows highly significant, large differences, $t(269) = 10.897$, $p < .001$, $r = .55$ (large effect); mean difference = 0.19 (95% *CI* = 0.15, 0.22). This comparison reveals that high offers ($5-$10) receive significantly higher social appropriateness ratings than fairness ratings.

I am not aware of any previous studies comparing fairness and appropriateness perceptions of strategic decisions. While comparing results across studies can be tricky because of differences in designs, looking at results reported in Krupka and Weber (2013), Erkut et al. (2014) and Rustichini and Villeval (2014) suggests a pattern of similarities and differences between fairness and appropriateness judgments broadly similar to that reported here:

1) In particular, in Rustichini and Villeval (2014) participants judge offers in UG exceeding approx. 66–71%

[9] I pool data from the KW game treatment and from the Non-incentivized treatment both in Figure 3 and in subsequent tests.

https://doi.org/10.1017/S1930297500003958 Published online by Cambridge University Press

of the pie as unfair[10]; similarly, offers in dictator game exceeding approximately 69–71% of the pie are rated as unfair. *In contrast*, all dictator offers exceeding the 50–50 split are still judged as socially appropriate on average in Krupka and Weber (2013) and Erkut et al. (2014). I.e., high offers are unfair (Rustichini & Villeval), but appropriate (Krupka & Weber, Erkut et al.)

2) At the same time, there does not seem to be such a misalignment between these three papers when it comes to ratings of fairness and social appropriateness of low offers: in Rustichini and Villeval (2014) participants judge offers below approx. 20–23% of the surplus in UG and below approx. 19–22% in dictator game as unfair. *Quite similarly*, offering 30% of the surplus or less in dictator game is rated as inappropriate on average in Krupka and Weber (2013) and Erkut et al. (2014). I.e., low offers are both unfair and inappropriate.

The results in these three papers are thus qualitatively similar to results reported in this article, i.e., low offers score low on both appropriateness and fairness, while high offers are rated as more appropriate than fair. New experiments are needed to obtain a deeper explanation of these observations.

4 Conclusion

The KW game is a promising technique of capturing socially shared judgments of appropriateness, fairness and potentially many other aspects of behavior. I elicit impartial ratings of social appropriateness and fairness of all possible offers in UG. However, I show that incentive-compatibility of this elicitation mechanism in not important in some contexts, such as the UG. It remains an open question whether incentives would affect judgments of social appropriateness and fairness in different, e.g., more complex contexts. I also find that ratings of social appropriateness and fairness are similar for low offers, but not for high offers which are judged to be more appropriate than fair.

References

Andreoni, J., & Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, *70*, 737–753.

Bardsley, N., & Sausgruber, R. (2005). Conformity and reciprocity in public good provision. *Journal of Economic Psychology*, *26*, 664–681.

[10]Depending on the player position from which the offers are rated. I report ratings from the first phase ("session") of Rustichini and Villeval's experiment.

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

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, *90*, 166–193.

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

Camerer, C., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, *19*, 7–42.

Cappelen, A. W., Hole, A. D., Sørensen, E. Ø., & Tungodden, B. (2007). The pluralism of fairness ideals: An experimental approach. *American Economic Review*, *97*, 818–827.

Croson, R., Handy, F., & Shang, J. (2009). Keeping up with the Joneses: The relationship of perceived descriptive social norms, social information, and charitable giving. *Nonprofit Management and Leadership*, *19*, 467–489.

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, *12*, 583–610.

Erkut, H., Nosenzo, D., & Sefton, M. (2014). Identifying social norms using coordination games: Spectators vs. stakeholders. Discussion paper.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics, 114*, 817–868.

Gächter, S., Nosenzo, D., & Sefton, M. (2013). Peer effects in pro-social behavior: Social norms or social preferences. *Journal of the European Economic Association*, *11*, 548–573.

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum games. *Journal of Economic Behavior & Organization*, *3*, 367–388.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*, 383–451.

Houser, D., & Xiao, E. (2011). Classification of natural language messages using a coordination game. *Experimental Economics*, *14*, 1–14.

Krupka, E. L., Leider, S., & Jiang, M. (2014). A meeting of the minds: Contracts and social norms. Working paper.

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

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dicta-

tor game sharing vary? *Journal of the European Economic Association*, *11*, 495–524.

Leliveld, M. C., van Beest, I., van Dijk, E., & Tenbrunsel, A. E. (2009). Understanding the influence of outcome valence in bargaining: A study on fairness accessibility, norms, and behavior. *Journal of Experimental Social Psychology*, *45*, 505–514.

Reuben, E., & van Winden, F. (2010). Fairness perceptions and prosocial emotions in the power to take. *Journal of Economic Psychology*, *31*, 908–922.

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

Rustichini, A., & Villeval, M. C. (2014). Moral hypocrisy, power and social preferences. *Journal of Economic Behavior & Organization*, *107*, 10–24.

van Winden, F. (2007). Affect fairness in economics. *Social Justice Research*, *20*, 35–52.

Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences*, *102*, 7398–7401.

Zelmer, J. (2003). Linear public goods experiments: A meta-analysis. *Experimental Economics*, *6*, 299–310.