OXFORD

Social Cognitive and Affective Neuroscience, 2016, 1608–1617

doi: 10.1093/scan/nsw069
Advance Access Publication Date: 3 June 2016
Original article

Behavioral and neuronal determinants of negative reciprocity in the ultimatum game

Laura Kaltwasser[1,2], Andrea Hildebrandt[3], Oliver Wilhelm[4], and Werner Sommer[1,2]

[1]Humboldt-Universität zu Berlin, Institut für Psychologie, Berlin 10099, Germany, [2]Berlin School of Mind & Brain, Berlin 10099, Germany, [3]Ernst-Moritz-Arndt-Universität Greifswald, Institut für Psychologie, Greifswald 17479, Germany and [4]Universität Ulm, Institut für Psychologie & Pädagogik, Ulm 89069, Germany

Correspondence should be addressed to Laura Kaltwasser, Humboldt-Universität zu Berlin, Berlin School of Mind and Brain, Unter Den Linden 6, Berlin 10099, Germany. E-mail: Laura.Kaltwasser@hu-berlin.de

Abstract

The rejection of unfair offers in the ultimatum game (UG) indicates negative reciprocity. The model of strong reciprocity claims that negative reciprocity reflects prosociality because the rejecting individual is sacrificing resources in order to punish unfair behavior. However, a recent study found that the rejection rate of unfair offers is linked to assertiveness (status defense model). To pursue the question what drives negative reciprocity, the present study investigated individual differences in the rejection of unfair offers along with their behavioral and neuronal determinants. We measured fairness preferences and event-related potentials (ERP) in 200 healthy participants playing a computerized version of the UG with pictures of unfair and fair proposers. Structural equation modeling (SEM) on the behavioral data corroborated both the strong reciprocity and the status defense models of human cooperation: Not only more prosocial but also more assertive individuals were more likely to show negative reciprocity by rejecting unfair offers. Experimental ERP results confirmed the feedback negativity (FN) as a neural signature of fairness processing. Multilevel SEM of brain–behavior relationships revealed that negative reciprocity was significantly associated with individual differences in FN amplitudes in response to proposers. Our results confirm stable individual differences in fairness processing at the behavioral and neuronal level.

Key words: ultimatum game; individual differences; fairness; social preferences; feedback negativity

Introduction

Can altruism, the thinking and acting in the interest of others (*alter* in Latin), include the motivation to harm another who does not cooperate? The concept of altruistic punishment describes acts of punishing non-cooperative behavior, which are costly for the actor and yield no material gains. Fehr and Gächter (2002) used a public good game in different rule contexts and showed that cooperation flourishes if altruistic punishment is possible but breaks down if punishment is ruled out. Individuals who commit altruistic punishment—the strong reciprocators—are willing to punish unfair behavior in others even though this is costly and does not provide any material rewards (Fehr *et al.*, 2002). According to the *strong reciprocity model* of the evolution of human cooperation, this form of negative reciprocity ensures the cooperation of future generations and therefore is a prosocial act (Axelrod and Hamilton, 1981).

The ultimatum game (UG) is a commonly used paradigm to study negative reciprocity. It is a two-stage game where two individuals, a proposer and a responder, bargain over a fixed amount of money. In the first stage, the proposer offers a split of his endowment; in the second stage, the responder choses to accept or reject the offer. If accepted, each player receives the split offered; if rejected, no player receives any money. By rejecting unfair offers, the responder shows negative reciprocity,

Received: 25 September 2015; **Revised:** 18 April 2016; **Accepted:** 13 May 2016

defying the rational solution according to economic theory (Nowak *et al.*, 2000).

Previous research suggests individual differences in negative reciprocity. For example, about 50% of the responders reject unfair offers if they receive <30% of the total sum (Camerer, 2003). But are these decisions to reject unfair offers really prosocial? Using the UG, Yamagishi *et al.* (2012) provided evidence against the strong reciprocity account because they found the personality trait of assertiveness instead of prosocial behavior to predict the rejection rate (RR) of unfair offers. They suggested that assertive participants use a tacit strategy to avoid the imposition of an inferior status. We termed this alternative explanation for negative reciprocity the *status defense model*. Clearly, people differ in their negative reciprocity proneness—but which motivation drives this inter-individual variance? There is a growing recognition that personality traits can help explain the heterogeneous responding within many economic games (Zhao and Smillie, 2015). Moreover, we believe that individual differences in physiological reactions to specific socio-economic interaction situations can help us to understand the underlying motivation.

Osinsky *et al.* (2014) studied the neural processes of the social evaluation process by recording the electroencephalogram (EEG) while participants played the UG, repeatedly receiving fair or unfair monetary offers from alleged other participants shown as portraits with neutral facial expressions. In that study, the faces could be used as predictive cues for the fairness of offers—two proposers each would always make fair or unfair offers, respectively. Osinsky *et al.* (2014) measured the feedback negativity (FN) in response to the portraits of the different proposers and to their offers. The FN is an event-related potential (ERP) consisting of a frontocentral negativity around 300–500 ms that is more pronounced after an unfavorable relative to a favorable event (Miltner *et al.*, 1997). It has been interpreted as an indicator of 'good *vs* bad evaluation' (Hajcak *et al.*, 2006) stemming from the dopaminergic signaling of reward prediction errors forwarded to medial frontal cortex (Gehring and Willoughby, 2002; Holroyd and Coles, 2002). The study by Osinsky *et al.* (2014) replicated the result that the FN was elicited by unfair relative to fair offers (Boksem and De Cremer, 2009; Hewig *et al.*, 2011; Van der Veen and Sahibdin, 2011; Wu *et al.*, 2011). Additionally, Osinsky *et al.* (2014) were able to show that over the course of the experiment, the FN was also elicited by the faces of unfair compared to fair bargaining partners, which suggests that the FN could reflect a special – social – instance of a more fundamental reward-prediction-error signaling originating from midbrain dopaminergic neurons (Sambrook and Goslin, 2015). This result confirmed previous research with functional magnetic resonance imaging where an affective value was ascribed to the opponent in repetitive interpersonal bargaining based on her/his fairness in the preceding interaction history (Singer *et al.*, 2004). Relative to neutral faces, faces of intentionally unfair or fair cooperators engendered increased activity in left amygdala, bilateral insula, fusiform gyrus, superior temporal sulcus and reward-related areas. Thus, Osinsky *et al.* (2014) revealed a basic neural mechanism of social evaluation during the UG, which is sensitive not only to the valence of monetary offers but also to learned fairness features of the proposers. This latter mechanism of social evaluation might also be indicative for individual differences in fairness preferences of the responders, that is, the evaluation of the proposers fairness may depend on the fairness preferences of the responder.

In the present study, we aimed to investigate individual differences in and neuronal and behavioral determinants of negative reciprocity in the UG with the particular intention to elucidate the triggering of the rejection of unfair offers. We aimed to establish brain–behavior relationships in order to explain which neural mechanisms might be related with inter-individual variance in negative reciprocity. Therefore, we applied a multilevel and multivariate approach using measures of personality and fairness preferences that are relevant for negative reciprocity next to the UG paradigm during EEG recording. In order to test the relative contribution of *strong reciprocity* and *status defense model*, we used the same measure of prosocial behavior [social value orientation (SVO); Murphy *et al.*, 2011) and a very similar questionnaire of assertiveness, as applied by Yamagishi *et al.* (2012), as predictors for negative reciprocity.

We also investigated the effect of the personality dimension honesty–humility (HH) of the HEXACO-PI-R (Lee and Ashton, 2004, 2006) on negative reciprocity. HH is thought to measure reciprocal altruism, namely the tendency toward active cooperation. Ashton and Lee (2007, p. 156) defined active cooperation as the tendency to be fair and genuine in dealing with others even if one might exploit them, whereas reactive cooperation was described as the tendency to be forgiving and tolerant of others, even if one might be exploited by them. Hilbig *et al.* (2013) found that HH predicted active cooperation in the dictator game, a variant of the UG, where the proposer simply states what the split will be and the responder has no veto power, but HH was not linked to reactive cooperation in terms of negative reciprocity in the UG.

Based on the aforementioned theory of strong reciprocity and the alternative explanation by Yamagishi *et al.* (2012), we hypothesized that both prosociality (strong reciprocity model) and assertiveness (status defense model) are linked to the rejection of unfair offers in UG. In contrast, HH should not predict the rejection of unfair offers in UG because it reflects active rather than re-active cooperation but share variance with measures of active cooperation such as prosociality. Following up on the work by Osinsky *et al.* (2014), we expected the FN to indicate the fairness of both, offers and proposers (pictured by their faces), specifically with unfair offers/proposers being processed more negatively. Moreover, we hypothesized that negative reciprocity impacts individual differences in FN in response to the fairness of the proposer. Our rationale was that participants with stronger fairness concerns in terms of negative reciprocity should show more pronounced fairness effects in FN in response to the proposer (face). In a neural status defense model, assertiveness should modulate this relative fairness effect in FN, whereas in a neural strong reciprocity model prosociality should account for it. Therefore, we ran a series of separate brain–behavior models to test the influence of each measure of fairness preferences on the fairness effect in FN.

Methods

Participants

The sample consisted of 210 healthy participants ($M_{age}=27.7$, $s.d._{age}=5.4$; 99 females) recruited from the participant pool of the Humboldt-Universität zu Berlin and newspaper advertisements. Participants received a compensation of 8 € per hour and were informed that they could win more money during UG depending on their choices. Each participant received an additional amount of 5 € as payout from UG. Ten participants were excluded from analyses due to low number of trials without artifacts (<15 trials per condition per block). The results after dropping these 10 participants (available upon request from the authors) did not differ from the ones described in the following.

All participants gave written informed consent; the study received ethics approval by the ethics committee of the Department of Psychology of the Humboldt-Universität zu Berlin.

Procedure

Behavioral session. The experiment was conducted in two sessions. During the behavioral session that lasted 2 hours, participants completed computerized self-report measures of personality and fairness preferences, described below, as well as several ability measures of face and object cognition, which are not the scope of this paper. All questionnaires were programmed in Inquisit software (Inquisit 4.0.0.1, 2012; Millisecond Software, Seattle, WA), and responses were given via computer mouse.

Assertiveness. We administered the assertiveness scale (selbstbehauptend) of the German *Inventory of Personality Styles and Disorders* (Persönlichkeits-Stil-und-Störungs-Inventar) (Kuhl and Kazén, 2009). The scale consists of 10 items ($\alpha = 0.82$) measuring the tendency to assert oneself when people attempt to impose their influence and the tendency to defend ones status. This tendency may extent to ruthless and antisocial behavior. A sample item is 'When someone goes against me, I can wear him/her down'. Responses are given on four-point Likert scales (disagree strongly, disagree somewhat, agree somewhat and agree strongly).

HH. Persons with high scores on the HH scale of HEXACO-PI-R (Lee and Ashton, 2004, 2006 for example items, see www.hexaco.org) avoid manipulating others for personal gain, feel little temptation to break rules, are uninterested in lavish wealth and luxuries, and feel no special entitlement to elevated social status. Conversely, persons with very low scores on this scale will flatter others to get what they want, are inclined to break rules for personal profit, are motivated by material gains, and feel a strong sense of self-importance. We used the 16 HH items of the 100-item version of HEXACO-PI-R ($\alpha = 0.91$). Participants responded on five-point Likert-type scales ranging from 'strongly disagree' to 'strongly agree'.

SVO. The concept SVO extends the rational self-interest postulate in economic theory by assuming that individuals tend to seek broader goals such as equality in outcomes. The magnitude of concern people have for others can be measured by a six-item questionnaire ($\alpha = 0.89$) about how participants would share resources with an anonymous stranger (Murphy *et al.*, 2011). Each item is a resource allocation over a continuum of joint payoffs. For instance, the participant has to choose a value x_{self} between 50 and 100, knowing that the anonymous partner will get $x_{other} = 150 - x_{self}$. According to the pay-off structure, the participant is assigned a continuous value of social orientation ($SVO = \arctan[(x_{other} - 50)/(x_{self} - 50)]$), which can be categorized to competitive, individualistic, prosocial and altruistic. Previous research indicates that SVO is a valid predictor of the cooperative tendency in social dilemmas (Bogaert *et al.*, 2008; Balliet *et al.*, 2009) and reflects the participants' true SVO rather than social desirability (Platow, 1994).

EEG session. The second session consisted of the UG during EEG recording. Upon arrival, participants were introduced to the rules of UG, informing them that they would play with other participants, which would require having their picture taken. Moreover, participants were asked to play the proposer in the UG, making 12 offers on a query sheet. In each offer, the participant could divide 10 cents into two shares: one for her/him and one for the other player. There were three predefined proposals: 9/1 (nine for the proposer, one for the responder), 7/3 and 5/5. Participants were informed that these offers would later be presented to other players who could then decide whether to accept or reject each offer. Participants were told that they would receive the corresponding amount of money if the offer was accepted by the responder. After providing their offers on a sheet, participants played the computerized version of the UG in the role of the responder while EEG was recorded. They were instructed that they would receive monetary offers made by six previous participants, but the actual offers came from six pseudo-proposers (50% females). These proposers were represented by portraits taken from a standardized stimulus set (Ebner *et al.*, 2010). We informed participants that the pictures of the proposers and their offers will be repeated several times in order to improve signal-to-noise ratio of the EEG. However, they were neither informed about the exact number of trials nor about the exact algorithm determining their additional payment depending on their performance in UG.

The UG comprised a total of 288 trials where all stimuli were presented at the center of a screen. Each trial started with a fixation cross shown for a variable duration of 500–1000 ms, followed by a photograph of a proposer for 1500 ms (Figure 1), and another fixation cross presented for 500–1000 ms; then, participants received an offer about splitting 10 cent presented in form of a pie chart and in written form. The offer could either be fair (5/5), slightly unfair (7/3) or highly unfair (9/1). By pressing the left or right response button, participants accepted or rejected the offer. After button press, a fixation cross was presented again for 500 ms. Participants received feedback about the sum booked to their account before the next trial started after 1250 ms.

The task was divided into three blocks (96 trials each), separated by self-timed breaks. Two of the proposers always made highly unfair offers (each 'unfair proposer' made 16 offers of a 9/1 split), two other proposers made fair offers (each 'fair proposer' made 16 offers of a 5/5 split) and two proposers made all kinds of offers with an equal frequency (each 'mixed proposer' made five offers of a 5/5 split, six offers of a 7/3 split and five offers of a 9/1 split). Thus, the relative offer probabilities were 0.44 for the 5/5 split, 0.44 for the 9/1 split and 0.11 for the 7/3 split. For each proposer type, one male and one female face were shown, with assignment of individual photographs to proposer categories being counterbalanced across participants. A fixed pseudo-random trial order was generated with the restriction that each proposer occurred twice within 12 trials, guaranteeing an equal distribution of single proposers across the task without immediate proposer repetition. After EEG recording participants rated each proposer's fairness on seven-point scales, ranging from very unfair (1) to very fair (7). Participants were debriefed after the experiment that they were actually playing against a computer.

The UG was programmed in Presentation 16.3 software (Neurobehavioral Systems Inc., Albany, CA, USA). During the task, participants sat in an acoustically and electrically shielded chamber on a comfortable chair with a distance of 60 cm between the head and the screen (17″). Each portrait was 18.8× 14.8 cm on the screen, resulting in a visual angle of 15.3°× 12.1°. The pie charts had a diameter of 9 cm (7.4° visual angle).

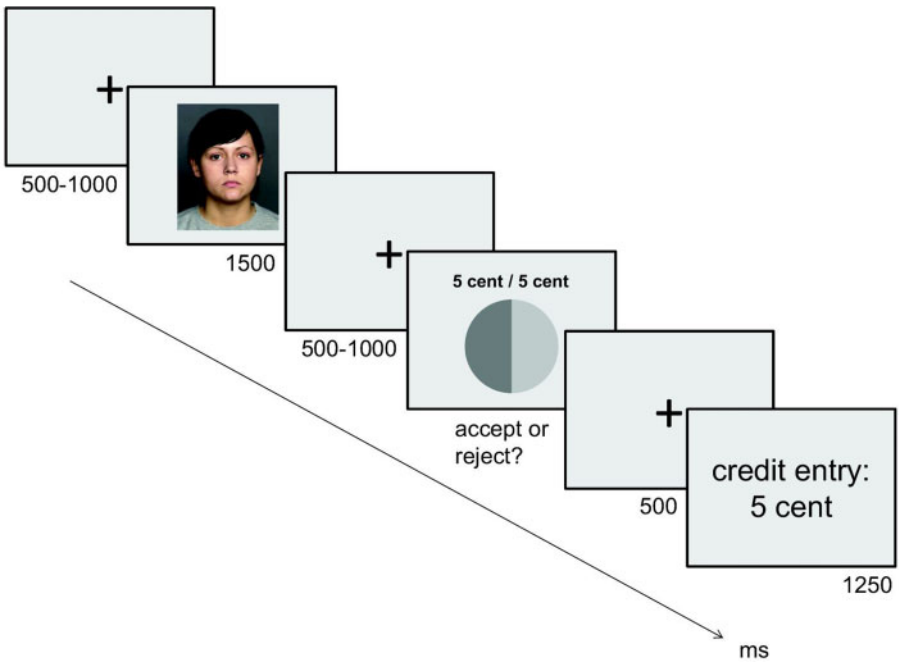

Fig. 1. Schematic depiction of a single trial in the UG.

Data recording and statistical analyses

EEG was recorded with 39 sintered Ag/AgCl electrodes mounted in an electrode cap (Easycap GmbH) and referenced to the left mastoid. Electrode AFz served as the ground. The horizontal electrooculogram (EOG) was recorded from two electrodes positioned at the outer canthi of the left and right eyes. The vertical EOG was monitored from Fp1, Fp2, and one additional electrode below the right eye. All channels were filtered with a band pass of 0.05–70 Hz and sampled at 1000 Hz. Impedances were kept below 5 kΩ. Offline, the continuous EEG signals were downsampled to 250 Hz, recalculated to an average reference, and segmented into 1200 ms epochs starting 200 ms before stimulus onset (separately for face and offer). These ERPs were digitally low-pass filtered at 20 Hz. Prior to segmentation, blinks, horizontal eye movements and pulse artifacts were removed by means of an automatic independent component analysis algorithm. Epochs with remaining artifacts were removed, according to the following criteria: maximal voltage difference within the epoch >150 μV and maximal voltage step of 20 μV/ms. At least 45 artifact-free trials were used for averaging per participant and condition. The ERPs were aligned to a 200 ms prestimulus baseline and averaged separately for each channel and experimental condition. The FN was measured as mean amplitude in the time window 220–352 ms following face onset and 300–400 ms following offer onset at electrodes F3, Fz, F4, FC1, FCz, FC2 and Cz. The time windows were determined by visual inspection of grand averages and global map dissimilarity (GMD; Lehmann and Skrandies, 1980), reflecting changes in topography over time as well as a recent meta-analysis on FN (Sambrook and Goslin 2015). The electrodes were chosen based on visual inspection of the ERPs and similar to previous studies on the FN in the UG (Osinsky *et al.*, 2014). All the reported processing steps up to here as well as single trial export (see below) were conducted in Brain Vision Analyzer 2.1 software.

Mean amplitudes of FN in the described time windows, RRs (percentage of rejected offers in relation to the total number of offers per offer type) as well as explicit ratings of fairness for each proposer face were analyzed with SPSS software (IBM, Armonk, NY, USA). For analyses including more than two levels repeated-measures analyses of variance (ANOVA) were conducted. In case of violation of sphericity assumption, epsilon (E; Greenhouse–Geisser correction) and corrected *P* values are reported. For pairwise comparisons, Bonferroni correction was applied.

Structural equation modeling. Latent factors of personality and fairness preferences and their mutual relationships were derived and validated in measurement models using confirmatory factor analysis with the *lavaan* package (Rosseel, 2012) in R system for statistical computing (R Development Core Team, 2008). Structural equation modeling (SEM) comprehensively tests theories about the linear relationships between multiple entities and explicitly accounts for measurement error (Bollen, 1989). The quality of a model is assessed by using multiple formal statistical tests and fit indices: χ^2 square statistic, root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR) and Comparative Fit Index (CFI). Moreover Akaike information criterion (AIC) and Bayesian information criterion (BIC) are used for model selection among a finite set of models.

Brain–behavior relationships. Single-trial amplitudes for within-person brain–behavior relationships of FN and fairness preferences were exported for electrode Cz, 220–352 ms following face onset. The single trial data were then fed into multilevel SEM (mSEMs) in order to test whether within-person modulations of FN amplitude by condition (level 1) are related to between-person differences in fairness preference (level 2). Commonly used in educational and developmental areas, multilevel models can also be useful for experimental designs with repeated measurements not involving time (Hoffman and Rovine, 2007). mSEMs were calculated with *Mplus 7* software (Muthén and Muthén, 1998–2012). Model comparisons for testing whether the

inclusion of variance components at level 2 increased model fit are based on Akaike information criterion (AIC).

Results

The results are separated into two sections: The within-person fairness effects, reporting the influence of proposer and offer type onto behavior and FN in the UG, and the between-person fairness effects, reporting the influence of personality traits onto behavior in the UG in terms of negative reciprocity. The between-person fairness effects are further explored by mSEMs, modeling the brain–behavior relationships of personality traits and negative reciprocity with FN.

Within-person fairness effects

RRs and explicit ratings. RRs and post-task fairness ratings were entered into a repeated-measures ANOVA with the within-subject factor 'proposer/offer type' (unfair, mixed and fair). Unfair offers (mean RR = 70.70%, SE = 2.61) were significantly more often rejected than mixed (mean RR = 30.64%, SE = 1.54) and fair offers (mean RR = 1.47%, SE=.34), $F(2,398)=574.88$, $P<0.001$, $\eta_p^2=0.74$.

Unfair proposers were rated as more unfair ($M=2.58$, SE = 0.09) than mixed ($M=4.15$, SE = 0.08) and fair proposers ($M=5.73$, SE = 0.08), $F(2,398)=309.38$, $P<0.001$, $\eta_p^2=0.61$. This shows that participants learned to discriminate between different proposer types.

Feedback-negativity. Face-locked FN was entered into a 7×3×3 repeated-measure ANOVA with the within-subject factors 'electrode' (F3, Fz, F4, FC1, FCz, FC2 and Cz), 'proposer type' (unfair, mixed and fair) and 'block' (first, second and third). Unfair and fair proposers elicited a FN as a relative negative deflection 220–352 ms after face onset compared to mixed proposers, $F(2,398)=15.26$, $P<0.001$, $\eta_p^2=0.07$ (Figure 2A). Post hoc pairwise comparisons revealed that unfair ($M=-2.31$, SE = 0.06) and fair proposers ($M=-2.31$, SE = .06) elicited a significantly more negative FN than mixed ones ($M=-2.04$, SE = .05), $P<0.001$. Moreover, there was a significant effect of block because participants showed a more pronounced FN in the first block ($M=-2.58$, SE = .18) than the second ($M=-2.04$, SE = .17) and third ($M=-2.02$, SE = .17) blocks ($F(2,398)=36.30$, $P<0.001$, $\eta_p^2=0.15$). However, there was no interaction between block and proposer type ($F(4,796)=4.74$, $p=0.552$).

The offer-locked FN was entered into a 7×3 repeated-measure ANOVA with the within-subject factors 'electrode' (F3, Fz, F4, FC1, FCz, FC2 and Cz) and 'offer type' (unfair, mixed and fair). There was a linear relationship between fairness of the offer and relative negativity of FN (Figure 2B) in that unfair

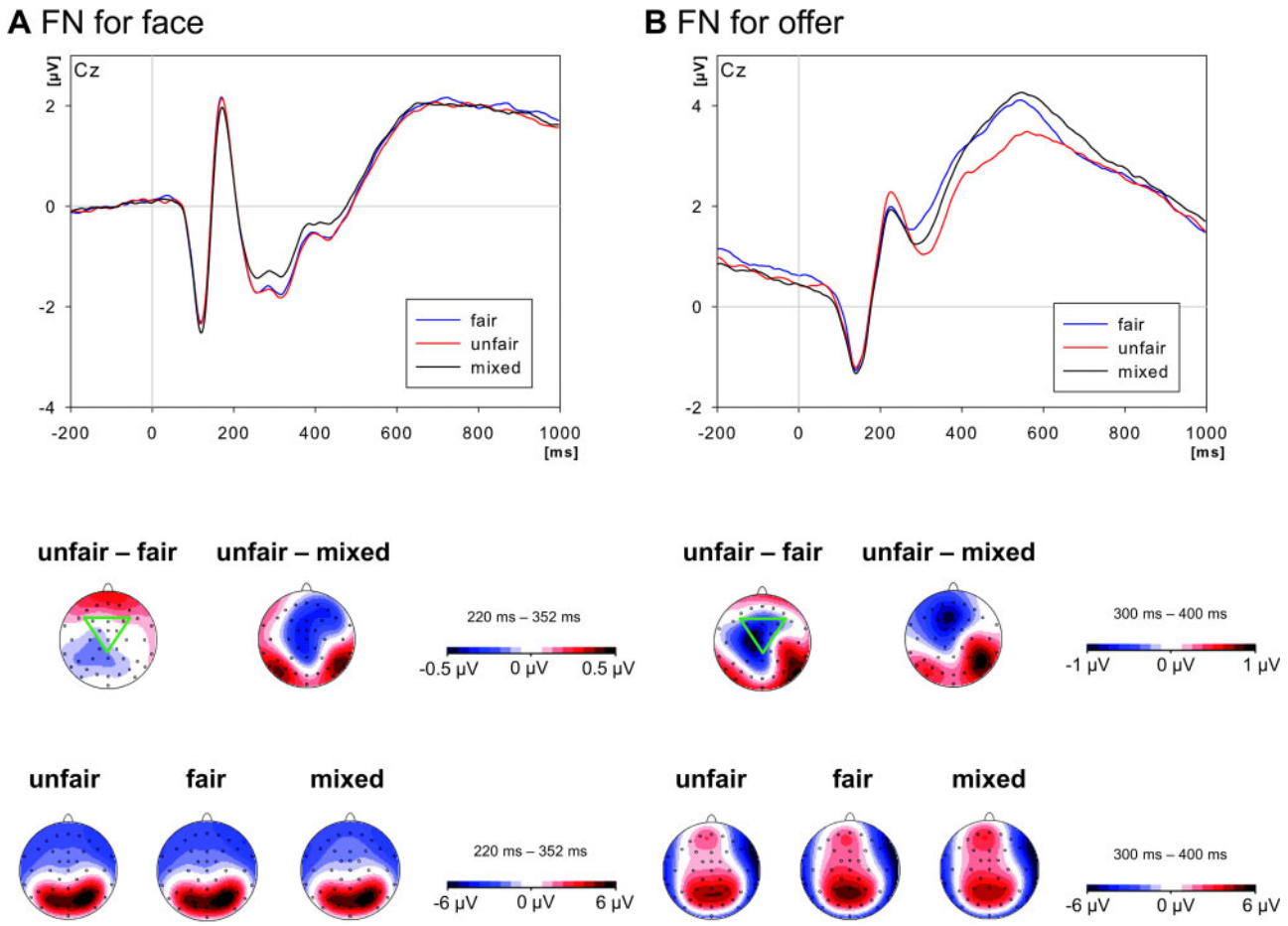

Fig. 2. ERPs locked to proposer face and offer in the UG. A, Face-locked FN at electrode Cz and scalp distribution of the unfair minus fair and unfair minus mixed differences (standard amplitude subtraction) in the time window 220–352 ms after proposer face onset. B, Offer-locked FN at electrode Cz and scalp distribution of the unfair minus fair and unfair minus mixed differences (standard amplitude subtraction) in the time window 300–400 ms after offer onset. Regions-of-interest for ANOVAs are indicated by green triangle on scalp.

offers elicited a significantly more negative FN than mixed and fair offers 300–400 ms after offer onset, $F(2,398) = 25.33$, $P < 0.001$, $\eta_p^2 = 0.11$. Post hoc pairwise comparisons showed that unfair ($M = 1.08$, $SE = 0.10$) differed significantly from mixed ($M = 1.69$, $SE = 0.11$) and fair offers ($M = 1.76$, $SE = 0.10$), $P < 0.001$.

Between-person fairness effects

SEM on fairness preferences. The measurement model of fairness preferences tested the influence of the latent factors assertiveness, HH, and prosociality on the latent factor negative reciprocity (Figure 3), for which the manifest RRs of unfair offers per block served as indicators. We formed three parcels as indicators for the latent assertiveness factor out of three items chosen with regard to item content. The indicators of the latent factor HH consisted of the mean scores of each subject on the subscales fairness, greed avoidance, modesty and sincerity. The latent prosociality factor comprised three parcels of two SVO items each and the total amount offered as a proposer in UG. The fit of this model was good ($\chi^2 = 89$, $df = 70$, $P = 0.067$, CFI = 0.981, RMSEA = 0.037, SRMR = 0.049). Both prosociality ($\beta = 0.35$, $P = 0.020$) and, to a lesser extent, assertiveness ($\beta = 0.24$, $P = 0.045$) significantly predicted the tendency to reject unfair offers (negative reciprocity). HH did not show a significant relationship ($\beta = 0.02$, $P = 0.919$). We performed two additional analyses in order to account for the non-normal distribution of the RRs: An additional model ($\chi^2 = 85$, $df = 70$, $P = 0.105$, CFI = .975, RMSEA = 0.034, SRMR = 0.046) with the MLR variant of maximum likelihood estimation using robust (Huber-White) SEs and a scaled test statistic that is (asymptotically) equal to the Yuan–Bentler test statistic decreased the influence of prosociality ($\beta = 0.35$, $P = 0.053$) and assertiveness ($\beta = 0.24$, $P = 0.080$) to a statistical trend, while the influence of HH remained unaffected. A Poisson regression with the RRs of unfair offers per block specified as count variables ($\chi^2 = 410$, $df = 970$, $P = 1.0$, CFI = 0.975, AIC = 8497, BIC = 8645) only revealed a significant influence of prosociality ($\beta = 0.21$, $P = 0.045$) on negative reciprocity, while the influence of HH ($\beta = 0.05$, $P = 0.612$) and assertiveness ($\beta = 0.01$, $P = 0.900$) were non-significant.

Brain–behavior relationships of fairness preferences. We considered two levels for the analyses of brain–behavior relationships of FN with fairness preferences in mSEMs. The first level tested the within-person experimental manipulation of the fairness of the proposer face (unfair, mixed and fair) eliciting the FN. The second level included between-person variations in the latent factors of fairness preferences and personality (negative reciprocity, prosociality, assertiveness and HH). Here, we examined whether the fairness condition effect in FN amplitude elicited by the face of the proposer was larger in participants with higher scores in negative reciprocity, prosociality, assertiveness or HH. We ran four separate models for each between-person latent variable on the second level. The indicators for each of these between-person latent variables were the same as in the SEM on fairness preferences (see above).

Level 1:

$$FN_{ij} = \beta_0 + \beta_1(C1_{ij}) + \beta_2(C2_{ij}) + e_{ij} \quad (1)$$

Level 2:

$$\beta_0 = \gamma_{00} + \gamma_{01}(\textit{negative reciprocity}) + u_{0i} \quad (2)$$

$$\beta_1 = \gamma_{10} + \gamma_{11}(\textit{negative reciprocity}) + u_{1i} \quad (3)$$

$$\beta_2 = \gamma_{20} + \gamma_{21}(\textit{negative reciprocity}) + u_{2i} \quad (4)$$

In the first level equation (1), the FN across persons i and trials j is described by an intercept (β_0), the average FN difference between mixed *vs* unfair and fair proposers across persons and trials (contrast code C1 [mixed = 2/3, unfair = −1/3, fair = −1/3],

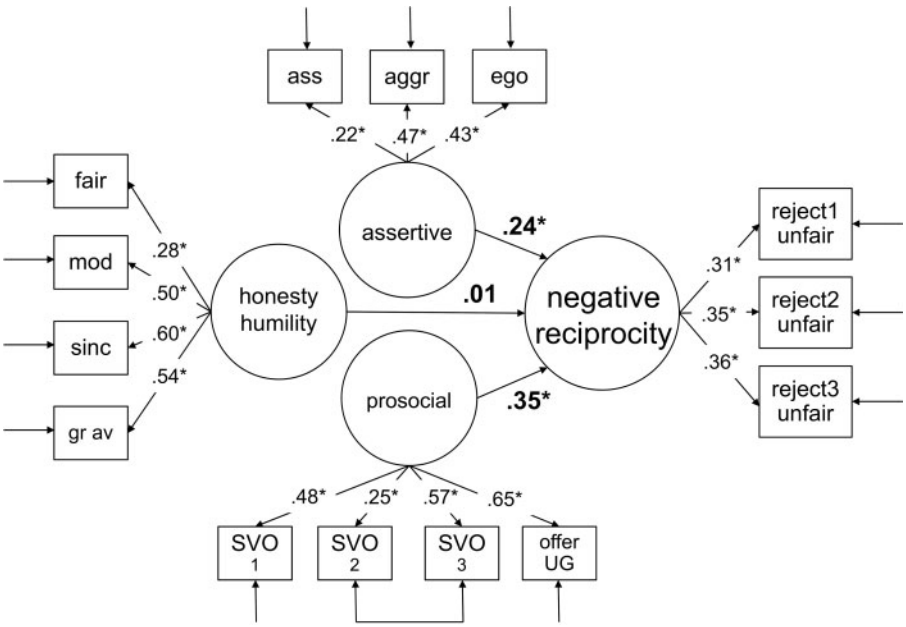

Fig. 3. SEM on fairness preferences. ($\chi^2 = 89$, df = 70, CFI = 0.981, RMSEA = 0.037, SRMR = 0.049). This is a schematic depiction. The correlations of the predictors were estimated.

with corresponding β_1), and the average FN difference between unfair *vs* fair proposers (contrast code C2 [mixed = 0, unfair = 1/2, fair = −1/2], with corresponding β_2). In the full model, these average fairness effects are allowed to vary across participants as the level 2 equations show. Note that as depicted in Table 1 model testing was carried out stepwise with the inclusion of random intercepts and random slopes at level 1 and subsequently the inclusion of between-person predictors at level 2. The modeled between-person variance components at level 2 (γ_{00}, γ_{10} and γ_{20}) around the fixed effects (β_1 and β_2) and their intercept (β_0) tested whether FN response to fairness conditions significantly differ across individuals. Their variation is predicted by the between-person level latent variable *negative reciprocity* (Figure 4). These prediction effects are reflected in the level 2 equations (2–4) by the parameters γ_{01}, γ_{11} and γ_{21}. The level 2 equations further depict between-person residual variances (u_{0i}, u_{1i} and u_{2i}), thus individual differences in the fairness effects on FN not predicted by *negative reciprocity* measured as an individual differences trait. Separate, but equivalent model series were estimated for each further level 2 predictor (prosociality, assertiveness, HH) of fairness effects on FN.

Table 1. Multilevel SEMs on the influence of fairness preferences and personality on FN amplitude differences between fairness conditions of proposers (mixed, fair and unfair)

	Model 1: negative reciprocity	Model 2: prosociality	Model 3: assertiveness	Model 4: HH
Level 1				
Intercept	−1.43 (0.18)	−1.43 (0.18)	−1.43 (0.18)	−1.43 (0.18)
C1 (mixed *vs* unfair & fair)	0.31* (0.06)	0.31* (0.06)	0.31* (0.06)	0.31* (0.06)
C2 (unfair *vs* fair)	−0.04 (0.07)	−0.04 (0.07)	−0.04 (0.07)	−0.04 (0.07)
Level 2				
Variances				
Intercept	6.07 (.62)	5.96 (.62)	6.06 (.62)	6.07 (.62)
C1	0.02 (0.06)	0.03 (0.06)	0.05 (0.06)	0.07 (0.07)
C2	0.20 (0.10)	0.20 (0.10)	.19 (0.10)	0.19 (0.10)
Predictor (γ parameters)				
Intercept	0.13 (.55)	−.79 (.51)	−0.22 (.46)	−0.05 (.37)
C1	0.34* (0.17)	−0.03 (0.15)	0.17 (0.15)	−0.05 (0.12)
C2	0.12 (0.21)	−0.03 (0.19	0.10 (0.18)	−0.15 (0.15)
Fit statistics				
df	19	22	19	22
AIC	343 167	346 310	344 136	345 458

df, number of free parameters.
*$p < 0.05$; SEs are depicted in parenthesis.

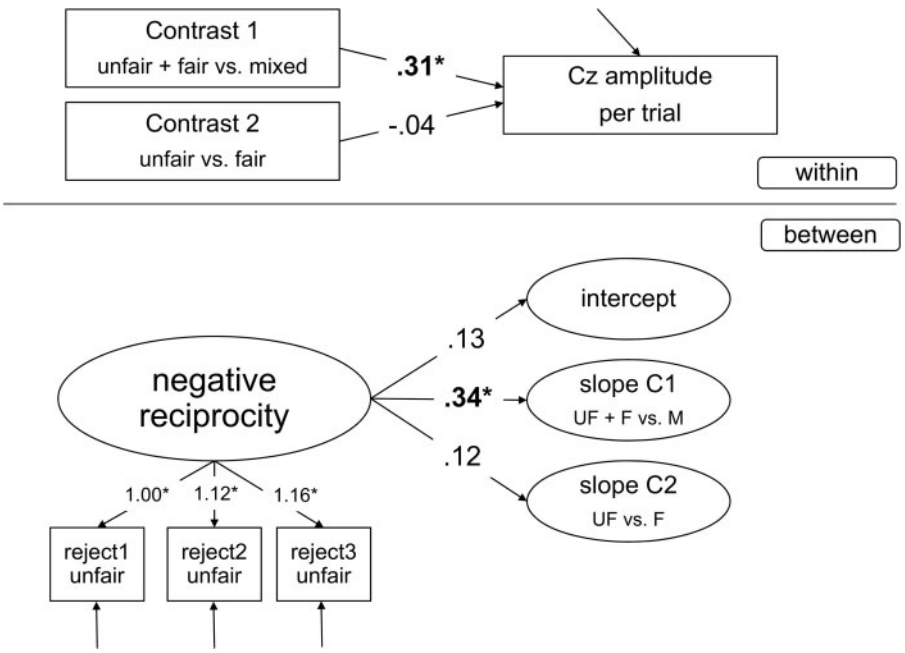

Fig. 4. Multilevel SEM on brain–behavior relationships of fairness preferences. The within-person level tests the influence of the experimental manipulation of the fairness of the proposer face on single trial FN at electrode Cz, comparing fair and unfair against mixed in contrast 1 (C1) and comparing unfair against fair in contrast 2 (C2). The between-person level tests whether the fluctuations in within-person brain–behavior relationships are related to individual differences in the latent factor of negative reciprocity.

At level 1, only C1 had a significant effect on FN amplitude ($\beta_1 = 0.31$, SE = 0.06, $P < 0.001$), indicating that mixed proposers elicited a less negative FN than unfair and fair proposers. As already expected from the within-person ANOVAs C2 did not show a significant effect ($\beta_1 = -0.04$, SE = 0.07, $P = 0.595$). AIC decreased (level 1: AIC = 343 251; level 2: AIC = 343 167) by including variance components for negative reciprocity at level 2 (γ_{00}, γ_{10} and γ_{20})—indicating better fit of this model as compared with the model estimating only fixed effects. In the four separate models with different latent between-person factors as predictors at level 2 (Table 1, models 1–4) only negative reciprocity significantly moderated the within-person fairness effect on FN amplitude ($\gamma_{11} = 0.34$, SE = 0.17, $P = 0.048$), indicating that participants with higher negative reciprocity also showed a stronger FN effect depending on the fairness of the proposer (Figure 5). Neither prosociality, assertiveness nor HH had a significant influence on the relative FN amplitude.

Because, arguably an influence of the proposer cannot be present before learning has occurred, we performed the reported mSEMs on the trials of the third block only. The pattern of results however barely changed as can be seen in Supplementary Table S1.

Discussion

Using a multivariate approach and abstracting from measurement error by means of SEMs on a large sample, we found that both prosociality (strong reciprocity model) and assertiveness (status defense model) predict negative reciprocity. Furthermore, experimental ERP results confirmed the FN as an indicator of social evaluation, reflecting fairness preferences toward the proposer in UG. A second step of analysis linked the experimental within-subject effects of fairness of the proposer on FN amplitude to the measurement model of individual differences in negative reciprocity. We used multilevel SEMs to investigate brain–behavior relationships of fairness preferences. The results revealed that the FN amplitude evoked by unfair and fair proposers relative to mixed ones was most pronounced in participants exhibiting stronger negative reciprocity in terms of RRs of unfair offers in UG. The mSEM did not provide evidence that assertiveness, HH or prosociality significantly explain fairness effects on the FN amplitude.

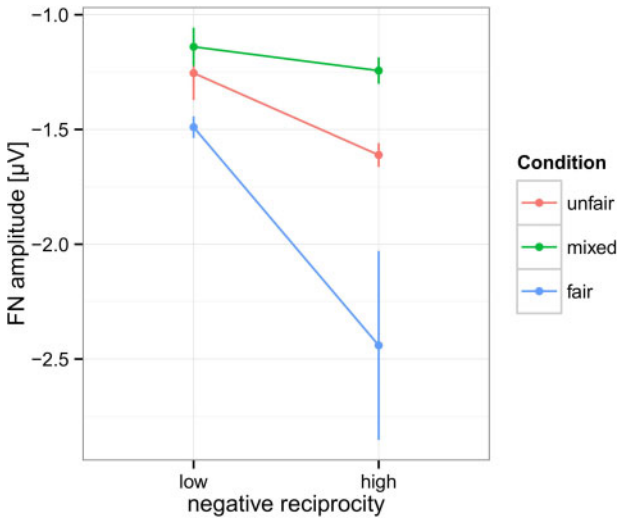

Fig. 5. Brain–behavior relationships of negative reciprocity for high *vs* low RRs using a median split (median = 0.79) on RR of unfair offers (± 95% confidence interval). Participants with high levels of negative reciprocity show a more pronounced fairness effect elicited by the proposer face in FN amplitude than participants with low levels of negative reciprocity.

Our results suggest several determinants of negative reciprocity, corroborating recent findings suggesting the existence of two types of rejecters in UG (Espín *et al.*, 2012; Branas-Garza *et al.*, 2014). On the one hand, our positive finding of prosociality predicting negative reciprocity supports the theory that prosocial participants inflict altruistic punishment and follow strong reciprocity by sacrificing their own resources in order to punish unfair behavior (Axelrod and Hamilton, 1981; Fehr *et al.*, 2002; Fehr and Gächter, 2002). This is in line with previous studies showing that prosocial individuals reciprocate by becoming uncooperative themselves if they are given clear evidence that the other person intentionally behaves uncooperatively (Kuhlman and Marshello, 1975; Van Lange, 1999). Especially in a repeated one-shot UG, where there is no opportunity to strategically punish unfair behavior, higher negative reciprocity of prosocial individuals can be interpreted as endorsing and enforcing a true norm of fairness. For example, an UG study by van Dijk *et al.* (2004) found that prosocial individuals made equal offers even when they had no apparent reason to fear that low offers would be rejected, while individualistic tended to make equal offers only when they had reason to fear that recipients would reject a low offer. On the other hand, our finding show that assertiveness also has a substantial relationship with negative reciprocity, suggesting that emotional styles or personality traits lead people to punish unfair behavior, allowing them to preserve integrity and avoiding the imposition of an inferior status (Yamagishi *et al.*, 2012). In situations of high interdependence, the need to defend ones status with negative reciprocity is reduced: Responders who were told that they would only be matched with another participant after UG, rejected an unfair offer more frequently than those in the interdependent condition who were told they had already been matched prior to making the decision (Declerck *et al.*, 2009). The authors explain the finding by unmatched responders rejecting more in order to signal that they are tough bargainers, fostering an illusion of control in terms of status defense.

On the neurophysiological level, our results confirm the FN as a mechanism of social evaluation during repeated interpersonal bargaining. As in previous studies, the FN in response to the offer showed a clear fairness modulation in that unfair offers elicited a larger (more negative) FN compared to mixed and fair offers characterized by a fronto-central negativity 300–400 ms after offer onset (Boksem and De Cremer, 2009; Hewig *et al.*, 2011; Van der Veen and Sahibdin, 2011; Wu *et al.*, 2011).

Furthermore, this effect transferred to the FN in response to the face of the proposer; portraits of unfair proposers elicited a relatively more negative FN, which partially replicates the results of Osinsky *et al.* (2014). Interestingly, we did not find a significant difference between unfair and fair proposers, but between both unfair and fair as compared to mixed proposers; this may suggest differential neural mechanisms being involved in the processing of fairness in faces compared to offers in our study. The FN found here in response to proposers seems to be an emotion signal coding a general social arousal or salience effect which might be similar to the processing of emotional context information in faces (Abdel Rahman, 2011; Wieser and Brosch, 2012). Recent experiments, examining the activity of single neurons of rhesus macaque monkeys, revealed distinct populations of dopaminergic midbrain neurons signaling motivational salience (Bromberg-Martin and Hikosaka, 2009; Bromberg-Martin *et al.*, 2010). We consider it plausible that the

increases in FN in response to fair and unfair faces reflect some kind of salience/alerting signal ('Pay attention! Something important is about to happen'.). Our brain–behavior analysis with multilevel SEM suggests that this salience effect is particularly evolved in strong reciprocators corroborating the results by Boksem and De Cremer (2009) where people with higher fairness concerns in terms of moral identity showed a more pronounced relative FN in response to offers in the UG. Using a dictator game with third-party punishment, Sun *et al.* (2015) analyzed the FN in response to unfair dictator offers in high and low altruists who could use their own endowment in order to punish the dictator. Surprisingly, they found opposite FN patterns in high and low altruists, reflecting different fairness considerations in those two groups. For high altruists, high unfair offers elicited a larger FN than medium unfair offers and fair offers. By contrast, for low altruists, fair offers elicited larger FN while high unfair offers caused the minimal FN. In our study, this differential neural fairness effect evoked by offers was transferred to the face of the proposer, suggesting that the FN is also a social signal reflecting social evaluation and processing of personal reputation. However, because the only other study examining the FN in response to the proposer in the UG obtained slightly different results, we emphasize that the evidence for FN as a signature of social preferences is still preliminary.

A limitation of the current study is the large number of choices each involving trivial incentives. This is however a general problem of many neuroeconomic studies trying to increase power by uprating the number of trials. Although we generally assume that low stakes due to minimal amounts of money (at the level of cents) may bias economic behavior in laboratory assessments as compared with real world socio-economic decisions, we do not assume that this limitation has serious implications for our research question. Importantly, our research aim was to investigate the processing of fairness based on proposer faces. Even though the overall manipulation of monetary stakes was not strong in the present study, the manipulation of the relative fairness of the proposer types had clear consequences, as manifested in the explicit behavioral effects at the level of RRs.

A second limitation concerns the fact that the association between prosocial behavior and negative reciprocity may be partly due to common method variance, since the SVO measure of prosocial behavior resembles a decomposed economic game similar to the UG, whereas assertiveness was measured by self-report. Future research should integrate more valid measures of personality traits. For example, Espín *et al.* (2012) operationalized impatience by an economic discounting task.

A third limitation regards the adequateness of our latent modeling techniques. Standard estimates of SEM presuppose normally distributed data, which is often not the case for RRs. Therefore, we also report more robust estimators in SEM (MLR) and alternative models applying RRs as count variables (Poisson regression). These analyses revealed less stable influences of assertiveness on negative reciprocity than of prosociality.

To conclude, our findings imply that one's own fairness considerations determine how we neurophysiologically process the social behavior of others. Our results were obtained with a sample size ($N = 200$) large enough to consider them reliable (Button *et al.*, 2013; Mar *et al.*, 2013). Future research should disentangle the origin of these idiosyncratic fairness preferences. Promising research already suggests that basal brain regions that are involved in reward and risk prediction are also recruited during the formation of these fairness preferences in the UG (Xiang *et al.*, 2013).

Acknowledgements

We thank Thomas Pinkpank, Rainer Kniesche and Hadiseh Nowparast for their valuable advice in data collection and analysis as well as Lena Fliedner, Alf Mante, Karsten Manske, Astrid Kiy, Friederike Rüffer, Tsvetina Dimitrova, Danyal Ansari, Katariina Mankinen, Susanne Stoll and Nina Mader for their help in recruitment and data collection.

Funding

This work was supported by a scholarship of Studienstiftung des deutschen Volkes to Laura Kaltwasser and a grant from the German Research Foundation to Andrea Hildebrandt [grant number HI 1780/2-1] and Werner Sommer [grant number SO 177/26-1].

Supplementary data

Supplementary data are available at *SCAN* online.

Conflict of interest. None declared.

References

Abdel Rahman, R. (2011). Facing good and evil: early brain signatures of affective biographical knowledge in face recognition. *Emotion*, **11**(6), 1397–405.

Ashton, M.C., Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review*, **11**(2), 150–66.

Axelrod, R., Hamilton, W. (1981). The evolution of cooperation. *Science*, **211**(4489), 1390–6.

Balliet, D., Parks, C., Joireman, J. (2009). Social value orientation and cooperation in social dilemmas: a meta-analysis. *Group Processes & Intergroup Relations*, **12**(4), 533–47.

Bogaert, S., Boone, C., Declerck, C. (2008). Social value orientation and cooperation in social dilemmas: A review and conceptual model. *British Journal of Social Psychology*, **47**(3),453–80.

Boksem, M.A.S., De Cremer, D. (2009). Fairness concerns predict medial frontal negativity amplitude in ultimatum bargaining. *Social Neuroscience*, **5**(1), 118–28.

Bollen, K.A. (1989). *Structural Equations with Latent Variables*, New York: Wiley.

Branas-Garza, P., Espin, A.M., Exadaktylos, F., Herrmann, B. (2014). Fair and unfair punishers coexist in the Ultimatum Game. *Scientific Reports*, **4**, 6025.

Bromberg-Martin, E.S., Hikosaka, O. (2009). Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron*, **63**(1), 119–26.

Bromberg-Martin, E.S., Matsumoto, M., Hikosaka, O. (2010). Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron*, **68**(5), 815–34.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., *et al.* (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews: Neuroscience*, **14**(5), 365–76.

Camerer, C.F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton: Princeton University Press.

Declerck, C.H., Kiyonari, T., Boone, C. (2009). Why do responders reject unequal offers in the Ultimatum Game? An experimental study on the role of perceiving interdependence. *Journal of Economic Psychology*, **30**(3), 335–43.

Ebner, N.C., Riediger, M., Lindenberger, U. (2010). FACES-A database of facial expressions in young, middle-aged, and older women and men: development and validation. *Behavior Research Methods*, **42**(1), 351–62.

Espín, A.M., Brañas-Garza, P., Herrmann, B., Gamella, J.F. (2012). Patient and impatient punishers of free-riders. *Proceedings of the Royal Society of London B: Biological Sciences*, *279*(1749), 4923–8.

Fehr, E., Fischbacher, U., Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, **13**(1), 1–25.

Fehr, E., Gächter, S. (2002). Altruistic punishment in humans. *Nature*, **415**(6868), 137–40.

Gehring, W.J., Willoughby, A.R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, **295**(5563), 2279–82.

Hajcak, G., Moser, J.S., Holroyd, C.B., Simons, R.F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biological Psychology*, **71**(2), 148–54.

Hewig, J., Kretschmer, N., Trippe, R.H., *et al.* (2011). Why humans deviate from rational choice. *Psychophysiology*, **48**(4), 507–14.

Hilbig, B.E., Zettler, I., Leist, F., Heydasch, T. (2013). It takes two: Honesty–Humility and Agreeableness differentially predict active versus reactive cooperation. *Personality and Individual Differences*, **54**(5), 598–603.

Hoffman, L., Rovine, M. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods*, **39**(1), 101–17.

Holroyd, C.B., Coles, M.G.H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, **109**(4), 679–709.

Kuhl, J., Kazén, M. (2009). *Persönlichkeits-Stil-und-Störungs-Inventar (PSSI) [Inventory of Personality Styles and Disorders]*, Göttingen: Hogrefe.

Kuhlman, D.M., Marshello, A.F. (1975). Individual differences in game motivation as moderators of preprogrammed strategy effects in prisoner's dilemma. *Journal of Personality and Social Psychology*, **32**(5), 922–31.

Lee, K., Ashton, M.C. (2004). Psychometric properties of the HEXACO Personality Inventory. *Multivariate Behavioral Research*, **39**(2), 329–58.

Lee, K., Ashton, M.C. (2006). Further assessment of the HEXACO Personality Inventory: Two new facet scales and an observer report form. *Psychological Assessment*, **18**(2), 182–91.

Lehmann, D., Skrandies, W. (1980). Reference-free identification of components of checkerboard-evoked multichannel potential fields. *Electroencephalography and Clinical Neurophysiology*, **48**(6), 609–21.

Mar, R., Spreng, R.N., DeYoung, C. (2013). How to produce personality neuroscience research with high statistical power and low additional cost. *Cognitive, Affective, & Behavioral Neuroscience*, **13**(3), 674–85.

Miltner, W.H., Braun, C.H., Coles, M.G. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: evidence for a "generic" neural system for error detection. *Journal of Cognitive Neuroscience*, **9**(6), 788–98.

Murphy, R.O., Ackermann, K.A., Handgraaf, M.J.J. (2011). Measuring Social Value Orientation. *Judgment and Decision Making*, **6**(8), 771–81.

Muthén, L.K., Muthén, B.O. (1998–2012). *Mplus User's Guide*, 7th edn, Los Angeles, CA: Muthén & Muthén.

Nowak, M.A., Page, K.M., Sigmund, K. (2000). Fairness Versus Reason in the Ultimatum Game. *Science*, **289**(5485), 1773–5.

Osinsky, R., Mussel, P., Ohrlein, L., Hewig, J. (2014). A neural signature of the creation of social evaluation. *Social Cognitive and Affective Neuroscience*, **9**(6), 731–6.

Platow, M.J. (1994). An Evaluation of the Social Desirability of Prosocial Self—Other Allocation Choices. *The Journal of Social Psychology*, **134**(1), 61–8.

R Development Core Team (2008). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, **48**(2), 1–36. URL http://www.jstatsoft.org/v48/i02/.

Sambrook, T.D., Goslin, J. (2015). A neural reward prediction error revealed by a meta-analysis of ERPs using great grand averages. *Psychological Bulletin*, **141**(1), 213–35.

Singer, T., Kiebel, S.J., Winston, J.S., Dolan, R.J., Frith, C.D. (2004). Brain Responses to the Acquired Moral Status of Faces. *Neuron*, **41**(4), 653–62.

Sun, L., Tan, P., Cheng, Y., Chen, J., Qu, C. (2015). The effect of altruistic tendency on fairness in third-party punishment. *Frontiers in Psychology*, **6**.

Van der Veen, F., Sahibdin, P. (2011). Dissociation between medial frontal negativity and cardiac responses in the ultimatum game: effects of offer size and fairness. *Cognitive, Affective, & Behavioral Neuroscience*, **11**(4), 516–25.

van Dijk, E., De Cremer, D., Handgraaf, M.J.J. (2004). Social value orientations and the strategic use of fairness in ultimatum bargaining. *Journal of Experimental Social Psychology*, **40**(6), 697–707.

Van Lange, P.A.M. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology*, **77**(2), 337–49.

Wieser, M.J., Brosch, T. (2012). Faces in context: a review and systematization of contextual influences on affective face processing. *Frontiers in Psychology*, **3**.

Wu, Y., Zhou, Y., van Dijk, E., Leliveld, M.C., Zhou, X. (2011). Social comparison affects brain responses to fairness in asset division: an ERP study with the ultimatum game. *Frontiers in Human Neuroscience*, **5**, 131.

Xiang, T., Lohrenz, T., Montague, P.R. (2013). Computational substrates of norms and their violations during social exchange. *The Journal of Neuroscience*, **33**(3), 1099–108.

Yamagishi, T., Horita, Y., Mifune, N., *et al.* (2012). Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(50), 20364–8.

Zhao, K., Smillie, L.D. (2015). The role of interpersonal traits in social decision making: exploring sources of behavioral heterogeneity in economic games. *Personality and Social Psychology Review*, **19**(3), 277–302.