Social Cognitive and Affective Neuroscience, 2016, 863–876

doi: 10.1093/scan/nsw026
Advance Access Publication Date: 29 March 2016
Original article

Altruistic traits are predicted by neural responses to monetary outcomes for self *vs* charity

René San Martín,[1,2,3] Youngbin Kwak,[1,4] John M. Pearson,[1,5] Marty G. Woldorff,[1,2,5,6] and Scott A. Huettel[1,2,5,6]

[1]Center for Cognitive Neuroscience, Duke University, Durham, NC 27710, USA, [2]Department of Psychology and Neuroscience, Duke University, Durham, NC 27710, USA, [3]Centro de Neuroeconomía, Facultad de Economía y Empresa, Universidad Diego Portales, Santiago 8370076, Chile, [4]Department of Psychological and Brain Sciences, University of Massachusetts, Amherst, MA 01003, USA, [5]Department of Neurobiology, Duke University Medical Center, Durham, NC 27710, USA, and [6]Department of Psychiatry and Behavioral Sciences, Duke University Durham, NC 27710, USA

Correspondence should be addressed to Scott Huettel, Center for Cognitive Neuroscience, Duke University, Box 90999, Durham, NC 27708, USA. E-mail: scott.huettel@duke.edu.

Downloaded from http://scan.oxfordjournals.org/ at Universidad Diego Portales on May 31, 2016

Abstract

Human altruism is often expressed through charitable donation—supporting a cause that benefits others in society, at cost to oneself. The underlying mechanisms of this other-regarding behavior remain imperfectly understood. By recording event-related-potential (ERP) measures of brain activity from human participants during a social gambling task, we identified markers of differential responses to receipt of monetary outcomes for oneself *vs* for a charitable cause. We focused our ERP analyses on the frontocentral feedback-related negativity (FRN) and three subcomponents of the attention-related P300 (P3) brain wave: the frontocentral P2 and P3a and the parietal P3b. The FRN distinguished between gains and losses for both self and charity outcomes. Importantly, this effect of outcome valence was greater for self than charity for both groups and was independent of two altruism-related measures: participants' pre-declared intended donations and the actual donations resulting from their choices. In contrast, differences in P3 subcomponents for outcomes for self *vs* charity strongly predicted both of our laboratory measures of altruism—as well as self-reported engagement in real-life altruistic behaviors. These results indicate that individual differences in altruism are linked to individual differences in the relative deployment of attention (as indexed by the P3) toward outcomes affecting other people.

Key words: altruism; monetary outcomes; event-related potentials; feedback-related negativity; P300

Introduction

Altruistic acts increase the welfare of others at a personal cost (Batson and Shaw, 1991). In most animal species, altruistic acts are exclusively directed toward kin (Hamilton, 1964). In humans, however, altruism goes far beyond helping genetically related individuals (Fehr and Fischbacher, 2003; Nowak and Sigmund, 2005) and is not limited to interpersonal interactions. Indeed, humans often sacrifice personal welfare on behalf of societal causes. Charitable donations of one's personal financial assets are a clear manifestation of this unique feature of human altruism, and an increasing number of studies have started focusing on decisions involving charities in order to shed light on other-regarding behavior (for a review, see Mayr *et al.*, 2008).

Recently, Kwak *et al.* (2014) conducted a behavioral study using a social gambling task (SGT) in which participants could learn to increase payoffs for themselves and/or for a charity institution. The authors reported an association between altruistic traits and the learning bias toward charitable outcomes (*vs* self-outcomes). Here, we use event-related-potential (ERP) measures of electrical brain activity and a modified version of

Received: 15 March 2015; **Revised:** 14 January 2016; **Accepted:** 8 March 2016

© The Author (2016). Published by Oxford University Press. For Permissions, please email: journals.permissions@oup.com

the SGT to evaluate the neurocognitive processes that could underlie this link between pro-social behavior and reward learning. The first hypothesis that we considered was that this link may be associated with differences in the way that the brain extracts utility from outcomes for oneself *vs* for others. Indeed, a characteristic feature of charitable giving is that it activates ventral striatal regions (Moll *et al.*, 2006; Harbaugh *et al.*, 2007) that typically respond to primary rewards (Apicella *et al.*, 1997; Delgado, 2007; Delgado *et al.*, 2014) and are considered to be a central component in the circuit for utility computations in the brain (Knutson and Peterson, 2005; Rangel *et al.*, 2008). However, recent work has also linked pros-social tendencies to the structure and function of the temporoparietal junction (Tankersley *et al.*, 2007; Morishima *et al.*, 2012; Carter and Huettel, 2013), a brain region typically associated with higher-level cognitive processes, including directing attention to relevant outcomes. This suggests, as a second hypothesis, that attentional processes unrelated to utility calculations could be involved in the link between altruistic traits and biases in reward learning reported by Kwak *et al.* (2014).

In this study, we took advantage of the high temporal resolution of the ERP technique in order to distinguish processes associated with utility extraction from processes associated with attentional allocation. Indeed, previous studies in humans suggest that these two sorts of process are indexed by two different ERP components: the feedback-related negativity (FRN) and the P300 (P3), respectively (San Martín, 2012). The FRN is a frontocentral negative-going ERP component that peaks ∼250 ms following the presentation of outcome information in decision-making tasks; its amplitude in response to outcomes tends to be proportional to the difference between acquired and expected utility (Holroyd and Coles, 2002; San Martín *et al.*, 2010). Source modeling studies have indicated that the FRN is likely generated, at least in part, in the anterior cingulate cortex in the frontal cortex (Miltner *et al.*, 1997; Gehring and Willoughby, 2002; van Schie *et al.*, 2004; Yu and Zhou, 2009). In contrast, the neural sources of the P3 are less clear and specific, with sources probably distributed across different regions of the cortex. In the context of learning-guided decision-making tasks, the P3 is thought to reflect attentional allocation during feedback evaluation (Nieuwenhuis *et al.*, 2005; Nieuwenhuis, 2011; San Martín, 2012; San Martín *et al.*, 2013). The P3 is composed of at least two distinguishable subcomponents: the early (peak ∼350 ms), frontally distributed P3a, which is thought to reflect stimulus-driven attentional processes, and the late (peak ∼450 ms), parietally distributed P3b, which is thought to reflect the amount of attention that is devoted to stimulus-induced memory or context updating (Polich, 2007). In reward learning studies, it is also important to characterize activity that precedes the onset of the FRN and has been labeled as ‘P2’ or ‘P200’ (peak ∼180 ms); this activity has been shown to be specifically associated with higher arousal levels (Carretié *et al.*, 2001; Schutter *et al.*, 2004) and attention capturing by target stimuli (Potts *et al.*, 1996, 2006; Potts, 2004).

Previous ERP studies employing monetary gambling tasks have demonstrated that the FRN and P3 are elicited not only in response to one’s own gains and losses but also in response to outcomes for others. For example, Itagaki and Katayama (2008) found that when others’ gains led to losses for the observer, the FRN to others’ gambling outcomes presented a reversed-polarity potential, being more negative in response to others’ gains than in response to others’ losses. Fukushima and Hiraki (2009) found that the FRN was elicited when observing the outcomes of decisions made by human agents, but not elicited in response to outcomes of decisions made by a computer. Furthermore, they found that self-reported measures of empathy toward the human agent were positively associated with the magnitude of this FRN. These studies suggest both that the FRN indexes a utility-based computation and that those computations can be applied to outcomes for others. The P3, in contrast, has been linked to a more general process of attentional allocation according to motivational/affective salience; like the FRN, though, this component is not specific to reward and social information (Leng and Zhou, 2010; Ma *et al.*, 2011)

Here, we evaluated whether social and non-social reward learning modulate these rapid ERP components—as predicted based on prior work—and tested two corollary predictions about the relationship between brain responses and pro-social behavior. First, we tested whether the propensity for pro-social learning (i.e. learning that increases benefits for a charitable organization) would covary with individual differences in the neural responses to pro-social outcomes (i.e. comparing charity *vs* self). Second, we evaluated whether such neural responses for charity *vs* self could predict self-reported engagement in real-life altruistic behaviors across individuals. If altruistic tendencies come into play during the rapid evaluation of utility, we would expect to see an association between individual differences in altruism and the FRN. In contrast, if altruistic tendencies alter higher-level attention-sensitive cognitive updating processes, we would expect to see an association between individual differences in altruism and the longer-latency P3.

Materials and methods

Participants

Forty-two healthy, right-handed, adult volunteers (22 male) participated in this study [ages, 18–34 years; mean (*M*) = 22.0]. Participants were recruited through the Duke Center for Cognitive Neuroscience Research Participation website (https://ccn-participate.sona-systems.com/). They signed up after reading the following description of the experiment: ‘If you choose to participate, you will wear a cap with electrodes that records your brain activity while playing a game in which you can win money for yourself and for a charity institution. The experimental session will take ∼2 h. You will be financially compensated at $15 per hour, with an extra bonus depending on the points earned during the experimental session’.

On average participants received $19.21 as extra bonus [standard deviation (s.d.) = $8.15]. The research team made a separate donation (M = $10.05, s.d. = $5.37) to the charity institution selected by the participant (see below), according to the points that the participant collected for the charity during the experimental session. Participants gave written informed consent and all procedures were approved by the Duke University Health System Institutional Review Board. We excluded four participants from ERP data analysis due to technical difficulties during their experimental sessions, leaving a final sample of 38 participants (20 male).

Behavioral measures

First, the participants filled out the Helping Orientation Questionnaire (HOQ), a paper questionnaire that measures four different personal orientations that may govern helping-related behavior in real life: altruistic, receptive-giving, inner-sustaining and selfish orientations (Romer *et al.*, 1986). Importantly, participants filled out this questionnaire before

Downloaded from http://scan.oxfordjournals.org/ at Universidad Diego Portales on May 31, 2016

receiving instructions about the task, and before responding to any other questionnaire.

Second, participants read instructions describing the study paradigm and four local charity institutions: Central NC Chapter of the American Red Cross, Animal Protection Society of Durham, Durham Literacy Center, Easter Seals of North Carolina. They were asked to choose one charity that would benefit from their participation in the experiment.

Third, we asked each participant to declare an intended donation using a paper questionnaire with the following question: 'Suppose that you win $50.00 in total during the task. How much of that money would you like to keep and how much would you like to donate to the institution that you just chose?' Before answering this question, subjects were informed that their donation to the charity would be based on their performance in the task and not based on their answer to this question. Responses to this question were later used to calculate an intended-donation score for each participant (see below, 'Questionnaires and behavioral data analysis' section).

Stimuli and task

We employed a probabilistic decision-making task based on the SGT designed by Kwak *et al.* (2014). Participants sat in front of a computer screen and performed 800 trials over the course of a single experimental session divided into 10 six-minute blocks. They were told that each trial would start with the presentation of four 'decks', each labeled with a different symbol. They were informed that the four decks were probabilistically associated with different payoffs for self and charity; by learning those probabilistic relationships, they could maximize the earnings for self and/or for charity, based on their own preferences. Also, they were told that the probabilistic relationship between decks and gains/losses would remain constant during each 80-trial block, and that new decks labeled with new symbols would appear at the beginning of each block. Subjects were also informed that no information regarding the conversion from points to money would be provided until the end of the experiment. Before data collection, participants completed a 20-trial practice session using a set of decks different from the ones used during data collection.

The temporal sequence of the task as it unfolded over a single example trial is shown in Figure 1A. Each trial started with the presentation of a fixation cross and four decks, each one labeled with a different unfamiliar symbol (i.e. a Japanese Hiragana character). The positions of the decks were randomized across trials, and they remained on the screen until the participant made a selection. Notice that, even though the position of the decks changed from trial to trial, participants could keep track of the identity of each deck based on the specific symbol that served as its label. If no response was made within 1200 ms, the words 'no response' were presented on the screen and a new trial started. The participant chose one of the decks by pressing one out of four joystick buttons that were spatially matched with the options on the screen. The selection was highlighted by superimposing a semitransparent yellow square on the chosen symbol for a period jittered in duration length between 100 and 200 ms. After an interstimulus interval (ISI) jittered between 600 and 800 ms, the outcomes for self and for charity were presented sequentially. The specific sequence of these outcomes (self-first/charity-second *vs* charity-first/self-second) was randomized across trials. Stars and circles indicated whether the outcome was for self or for charity, and the specific symbol-recipient association (stars-self/circles-charity *vs* stars-charity/circles-self) was randomized across participants. Won points were presented as green symbols and lost points were presented as red symbols. For example, Figure 1A presents the case of losing three points for self and winning two points for charity (for a participant for whom stars presented outcomes for self and circles presented outcomes). The presentation of outcomes for self and charity were temporally separated by an ISI jittered between 600 and 800 ms, and the next trial started after an intertrial interval jittered between 800 and 1200 ms. Participants were instructed to maintain fixation on the central cross throughout the experimental runs.

The outcome for self and charity on each trial was probabilistically set according to the following formula. With regard to the valence (win *vs* loss) of the outcomes for self and charity, each deck was associated with a specific probability for each of four possible scenarios: Self-win\Charity-win, Self-win\Charity-loss, Self-loss\Charity-win and Self-loss\Charity-loss (Figure 1B). According to those probabilities, one of the decks tended to deliver gains both to self and charity (S+\C+), a second one tended to deliver even more frequent gains to self but losses to charity (S++\C−), a third deck tended to deliver losses to self but highly frequent gains to charity (S−\C++) and finally a fourth deck tended to deliver losses both to self and charity (S−\C−). With regard to the magnitude of the gains or losses for self and charity, the stimulation software randomly selected one out of four possible outcome magnitudes (1, 2, 3 or 4 points; 25% chances each), independently for self and charity. Importantly, participants' choices affected the probability of winning *vs* losing but not the number of points that were won or lost.

As a result, deck S+\C+ was associated with an expected value (EV) of 0.5 points for self and 0.5 points for charity, deck S++\C− with an EV of 1.25 points for self and −0.25 points for charity, deck S−\C++ with an EV of −0.25 points for self and 1.25 points for charity and deck S−\C− with an EV of −0.25 points both for self and for charity (Figure 1B).

Participants could evenly split earnings between self and charity by learning to select deck S+\C+ or by alternating between decks S++\C− and S−\C++. Alternatively, they could maximize earnings for self by learning to select deck S++\C+ (at the expense of earnings for charity) or could maximize the charity's earnings by learning to select deck S−\C++ (at the expense of their own earnings). Importantly, three decks (S+\C+, S++\C− and S−\C++) were equal in EV (i.e. $EV_{self}+EV_{charity}=1$, for these three decks), differing only on how those points were distributed between self and charity. Deck S−\C− had a negative EV both for self and for charity.

Analyses of behavior

We calculated an intended-donation score for each participant as the ratio between their declared intended donation and the total amount considered in the paper questionnaire ($50); this score was thereby normalized between the values of 0 and 1. In subsequent analyses in the results, we distinguished two groups of participants, an altruistic-group and a selfish-group, based on the distribution of their intended-donation scores. Note however that the label 'selfish-group' here is used in a relative sense, since even a small donation represents an altruistic act. We also calculated an actual-donation score for each participant as the ratio between the number of points earned during the task for the charity and the total number of points earned during the task $[points_{charity}/(points_{charity}+points_{self})]$. Finally, we also calculated the altruism subscore of the HOQ for each participant following the scoring method indicated by

Downloaded from http://scan.oxfordjournals.org/ at Universidad Diego Portales on May 31, 2016

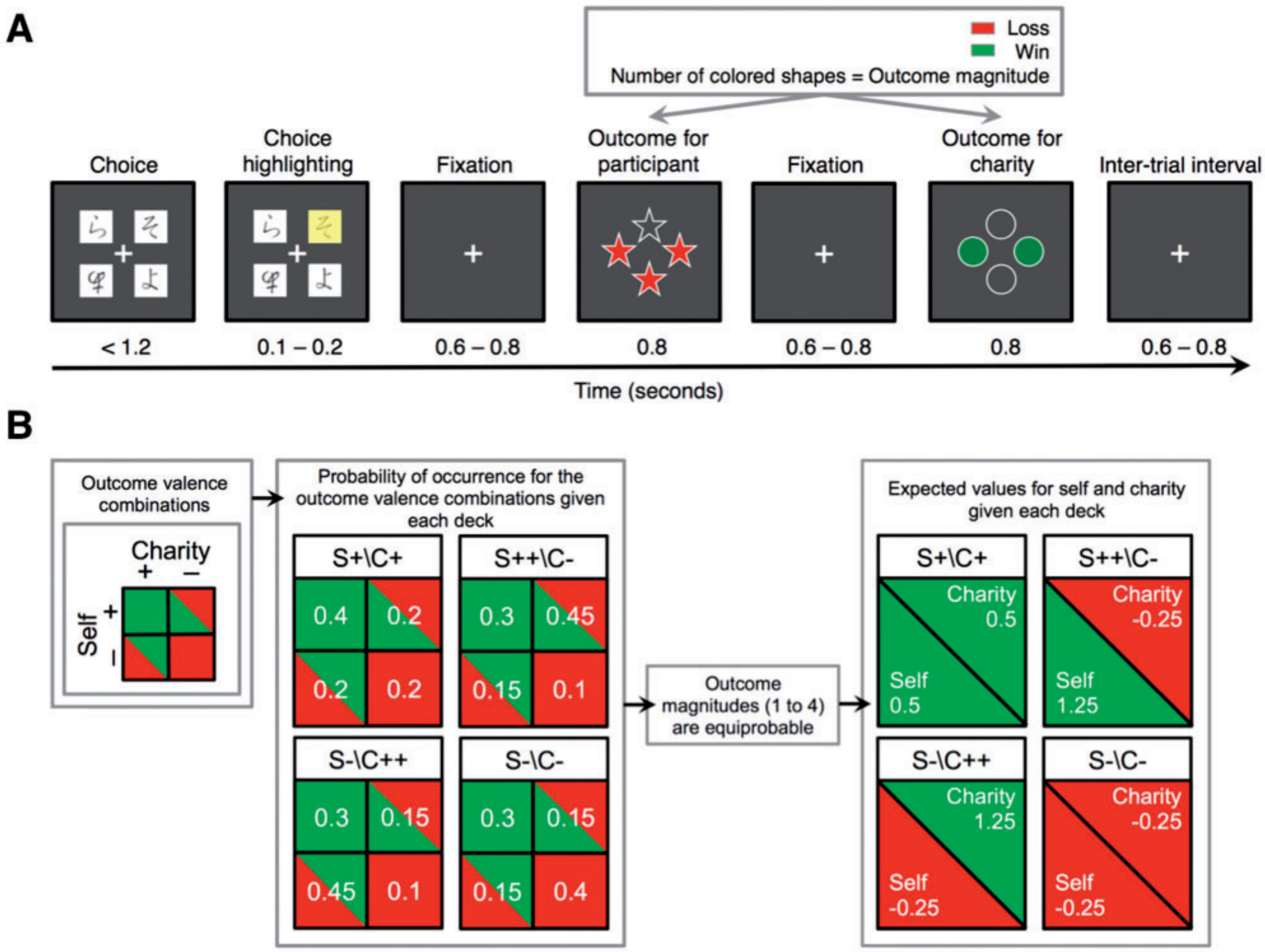

Fig. 1. Experimental design. **(A)** Participants chose between four decks, each one labeled with a different symbol. The position of these decks was randomized across trials, but the symbol label for each deck remaining constant from trial to trial. Feedback concerning outcomes for self and for the charity was sequentially presented as a number of colored stars and circles. The specific symbol-recipient association (stars-self/circles-charity *vs* stars-charity/circles-self) was randomized across participants (although consistent within a participant), and the specific sequence (self-first/charity-second *vs* charity-first/self-second) was randomized across trials for each participant. Red-colored shapes indicated the number of points lost and green-colored shapes indicated the number of points won. **(B)** There were four possible outcome valence combinations on each trial: Self-win\Charity-win, Self-win\Charity-loss, Self-loss\Charity-win and Self-loss\Charity-loss. The outcome on each trial was randomly determined according to the weight that the chosen deck assigned to each of these outcome valence combinations. For example, deck S++\C- was associated with a 30% chance of observing a self-win\charity-win outcome, 45% for self-win\charity-lose, 15% chance for self-lose\charity-win and 10% chances for self-lose\charity-lose. Outcome magnitudes (1–4) were equiprobable (i.e. 25% each), resulting in specific EVs for self and charity associated with each deck.

Romer *et al.* (1986). Using these values, we evaluated the participant-wise correlation between intended-donation, actual-donation and altruism scores.

We also evaluated whether participants in the altruistic-group and the selfish-group differed in the earnings that they collected during the experimental session. We included the participants' $\text{points}_{\text{self}}$ and $\text{points}_{\text{charity}}$ in an analysis of variance (ANOVA), along with the within-subjects factor recipient (self/charity) and the between-subject factor of group based on intended-split scores (altruistic-group/selfish-group). Tukey's honest significant difference (HSD) method was used in the calculation of *post hoc* contrasts throughout our behavioral and ERP data analyses.

Finally, we evaluated whether altruistic and selfish participants differ in the degree in which outcomes for self and charity impacted behavioral adjustments on a trial-by-trial basis. We considered each trial (*t*) in terms of the deck that was chosen on *t*. We then calculated the observed change in the probability for persevering in choosing a given deck across trials $t+1$, $t+2$ and $t+3$ as a function of whether the transitions between trials corresponded to gains or losses. We submitted this change in perseverance probability to an ANOVA with two within-subject factors [recipient (self/charity), valence (losses/gains)] and one between-subject factor of group based on intended-split scores (altruistic-group/selfish-group).

Electroencephalogram recording and preprocessing

The electroencephalogram (EEG) was recorded continuously from 64 active-electrode channels mounted in a customized, extended coverage, elastic cap (Brain Products ActiCap) using a bandpass filter of 0.01–100 Hz at a sampling rate of 500 Hz (Brain Products ActiChamp). All channels were referenced to the right mastoid during recording. The positions of the 64 channels were equally spaced across the customized cap and covered the whole head from slightly above the eyebrows to below the inion (Woldorff *et al.*, 2002). Impedances of all channels were kept below 15 kΩ, and fixation was monitored with horizontal and vertical electrooculogram recordings. Recordings took place in

Downloaded from http://scan.oxfordjournals.org/ at Universidad Diego Portales on May 31, 2016

an electrically shielded, sound-attenuated and dimly lit experimental chamber.

Offline, EEG data were exported to MATLAB (MathWorks) and processed using the EEGLAB software suite (Delorme and Makeig, 2004) and custom scripts. The data were low-pass filtered at 40 Hz using linear finite impulse response filtering, down-sampled to 250 Hz, and re-referenced to the algebraic average of the left and right mastoid electrodes. For each participant, we implemented a procedure for artifact removal based on an independent component analysis (ICA) approach that has been established previously (Debener *et al.*, 2005; Eichele *et al.*, 2005; Scheibe *et al.*, 2010; San Martín *et al.*, 2016) to obtain EEG data with greatly diminished contribution from ocular/biophysical artifacts. More specifically, we first visually rejected unsuitable portions of the continuous EEG data containing obvious non-neural noise artifacts (e.g. drifts, excessive blinks). Second, we separated the remaining data into 1200 ms feedback-locked epochs, spanning from 400 ms before to 800 ms after the onset of the feedback stimulus, using a prestimulus baseline period of 200 ms. Third, we performed a temporal infomax ICA (Bell and Sejnowski, 1995). Fourth, independent components with scalp topographies and signals that could be assigned to known artifacts (e.g. blinks, heart beats) were removed from the data (Jung *et al.*, 2000a,b; Delorme *et al.*, 2007). The remaining components were back-projected to scalp time-amplitude space to create an artifact-corrected EEG data set.

ERP data analysis

The ERP components analysed here include the FRN and three subcomponents of the P3 response: the P2 (early-P3a), P3a and P3b. The FRN has a frontocentral distribution with a typical peak of amplitude over the standard 10–20 FCz location at ~250 ms after feedback onset (Miltner *et al.*, 1997; Gehring and Willoughby, 2002; Nieuwenhuis *et al.*, 2004; San Martín *et al.*, 2010). The P3 is usually formed by two subcomponents: the P3a with a frontocentral distribution and a maximum amplitude between 300 and 400 ms following stimulus presentation, and the P3b with a parietocentral distribution and a peak of amplitude occurring ~100 ms later (Nieuwenhuis *et al.*, 2005; Polich, 2007; San Martín *et al.*, 2013). At frontal sites, there is an overlap between the FRN and the P3a, and studies have used the label 'P2' or 'P200' to refer to P3a activity that precedes the onset of the FRN (Rigoni *et al.*, 2010; San Martín *et al.*, 2010, 2013). The superimposition of the FRN on the P3a also poses a challenge to the assessment of the FRN since, as several studies have noted, the FRN peak can be shifted depending on the amplitude of this frontal P3a (Yeung and Sanfey, 2004; Chase *et al.*, 2011; Billeke *et al.*, 2012; San Martín *et al.*, 2013).

To establish an unbiased, a priori method for identifying ERP components of interest, we adopted both the regions of interest and time windows from a previous study (San Martín *et al.*, 2013). In particular, we defined a frontocentral region of interest (ROI) of seven sensors centered on the standard FCz channel, and a parietocentral ROI of seven sensors centered on the standard Pz channel. We used the frontocentral ROI to calculate the P2, FRN and P3a. For each trial, the P2 was calculated as the average ERP voltage potential from a 152 to 184 ms postfeedback window. (Note that the effective sampling rate was 250 Hz, and thus these window lengths were all multiples of 4 ms.) To more effectively quantify the FRN amplitude accounting for differences in the P3-induced baseline, we used a mean amplitude to mean amplitude approach. More specifically, the FRN amplitude was calculated as the average potential across a 204–272 ms window after feedback, minus the average voltage potential from a short 188–200 ms anchoring window preceding it. This subtraction was performed in order to diminish the impact of differences in the P3-induced baseline on the FRN measurement. The P3a was quantified as the average potential from a 284 to 412 ms window, also in the frontal ROI. The P3b was defined by the average potential from a 416 to 796 ms window, now using the parietal ROI.

Repeated-measures ANOVAs analysed the ERP components with four within-subject factors: recipient (self/charity), valence (loss/win), magnitude (1–4), behavioral adjustment (switch/stay), along with one between-subjects factor of group based on intended-split scores (altruistic-group/selfish-group). The inclusion of the factor behavioral adjustment was aimed at assessing the association between each ERP component and the behavioral adjustment on subsequent trials. In order to define levels of subsequent behavioral adjustment for the ERP data we considered each trial *t* in terms of the chosen deck, and we then asked whether the choice was the same or different on the next trial ($t+1$). If the choice on $t+1$ was the same than on trial *t*, we assigned the ERP data on trial *t* to a 'stay' bin. Otherwise, we assigned the ERP data on trial *t* to a 'switch' bin.

Results

Behavior

Before the task, participants were asked to declare what proportion of the money earned during the task they would be willing to donate to the charity. On average they declared themselves willing to donate 28.3% of that money (s.d. = 22.0%). The mean intended-donation score was numerically larger for females (M = 35%, s.d. = 23.8%) than males (M = 22.3%, s.d. = 18.7%), but this trend did not reach statistical significance ($t_{(36)}$=1.84, P =0.07).

The distribution of these intended-donation scores (Figure 2) revealed two clearly distinguishable groups of participants: a group of participants with intended-donation scores lower than 30% (*M* =10.3%, s.d. = 7.4%) and a second group with intended-donation scores higher than 30% (*M* = 50.6%, s.d. = 9.7%). Hereafter, we refer to these as the selfish-group (*n* = 21) and the altruistic-group (*n* = 17), respectively.

Across participants, intended-donation scores covaried both with actual-donations resulting from the experimental task ($r_{(36)}$=0.64, $P < 0.001$) and with the altruism subscores of the HOQ ($r_{(36)}$=0.40, P = 0.01), suggesting that participants' declared donation preferences greatly influenced the actual reward-earning results in the task and that these preferences were aligned with their self-reported engagement in real-life altruistic behaviors. Altruism HOQ subscores and actual-donations also tended to covary, but this correlation did not reach statistical significance ($r_{(36)}$=0.29, P = 0.08). There were no significant gender differences in the altruism subscores of the HOQ ($t_{(36)}$=0.66, *P* =0.51) nor in actual-donations ($t_{(36)}$=−0.19, P =0.85).

Next, we analysed the points that participants earned for each of the potential recipients (i.e. self and charity) using an ANOVA that included the between-subjects factor of group. This analysis revealed a main effect of recipient, showing that, in general, participants collected more points for self [M = 384.4 points (pts), s.d. = 160.1] than for the charity (M = 201.0 pts, s.d. = 100.1) [F(1, 36)=30.9, P< 0.001]. We also found an interaction between recipient and group [F(1, 36)=22.1, P< 0.001], such that the difference between points for self and points for charity was significantly greater in the selfish-group (M = 310.9

Downloaded from http://scan.oxfordjournals.org/ at Universidad Diego Portales on May 31, 2016

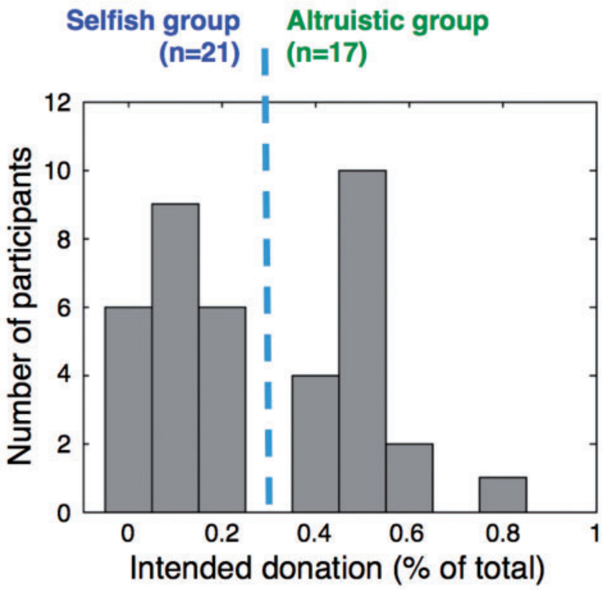

Fig. 2. Distribution of intended-donation scores. Participants were asked about the proportion of the money earned during the task they would be willing to donate to the charity. Based on the distribution of their responses we defined a selfish-group and an altruistic-group of participants. This group definition was used as a between-subjects factor in several of our subsequent behavioral and ERP data analyses.

pts, s.d. = 230.1) than in the altruistic-group (*M* = 25.8 pts, s.d. = 106.8).

Interestingly, we also found a main effect of group, showing that participants in the selfish-group collected more points (*M* = 633.5 pts, s.d. = 117.7) than participants in the altruistic-group (M = 525.9 pts, s.d. = 124.1), regardless of the distribution of those points between self and charity [F(1, 36)=7.5, $P < 0.01$]. Since decks $S+\backslash C+$, $S{+}{+}\backslash C-$ and $S-\backslash C{+}{+}$ were associated with the same total number of awarded points (i.e. $EV_{self} + EV_{charity} = 1$, for these three decks), the main effect of group indicates that participants in the selfish-group were better than participants in the altruistic-group in avoiding the selection of deck $S-\backslash C-$, which had a negative EV for both self and charity. In order to test the association between altruism and the probability of choosing $S-\backslash C-$ *vs* any of the other decks [i.e. $p(S-\backslash C-)$], we fitted a logistic regression model according to the following equation: $p(S-\backslash C-) = 1/(1 + e-^{z})$, where $z = \beta_0 + (\beta_1)$ * $(\gamma_{intended\_donations})$. In this equation '$\gamma_{intended\_donations}$' represents the intended-donation scores for each participant. This analysis confirmed that intended-donation scores were positively associated with choosing $S-\backslash C-$ ($\beta_1 = 0.8$, $P < 0.01$). In other words, the more altruistic participants were (as reflected by their intended-donation scores), the worse they were at avoiding deck $S-\backslash C-$.

We also evaluated the observed probability of persevering across trials as a function of gains or losses to each recipient. A visual inspection of Figure 3 suggests that participants in the

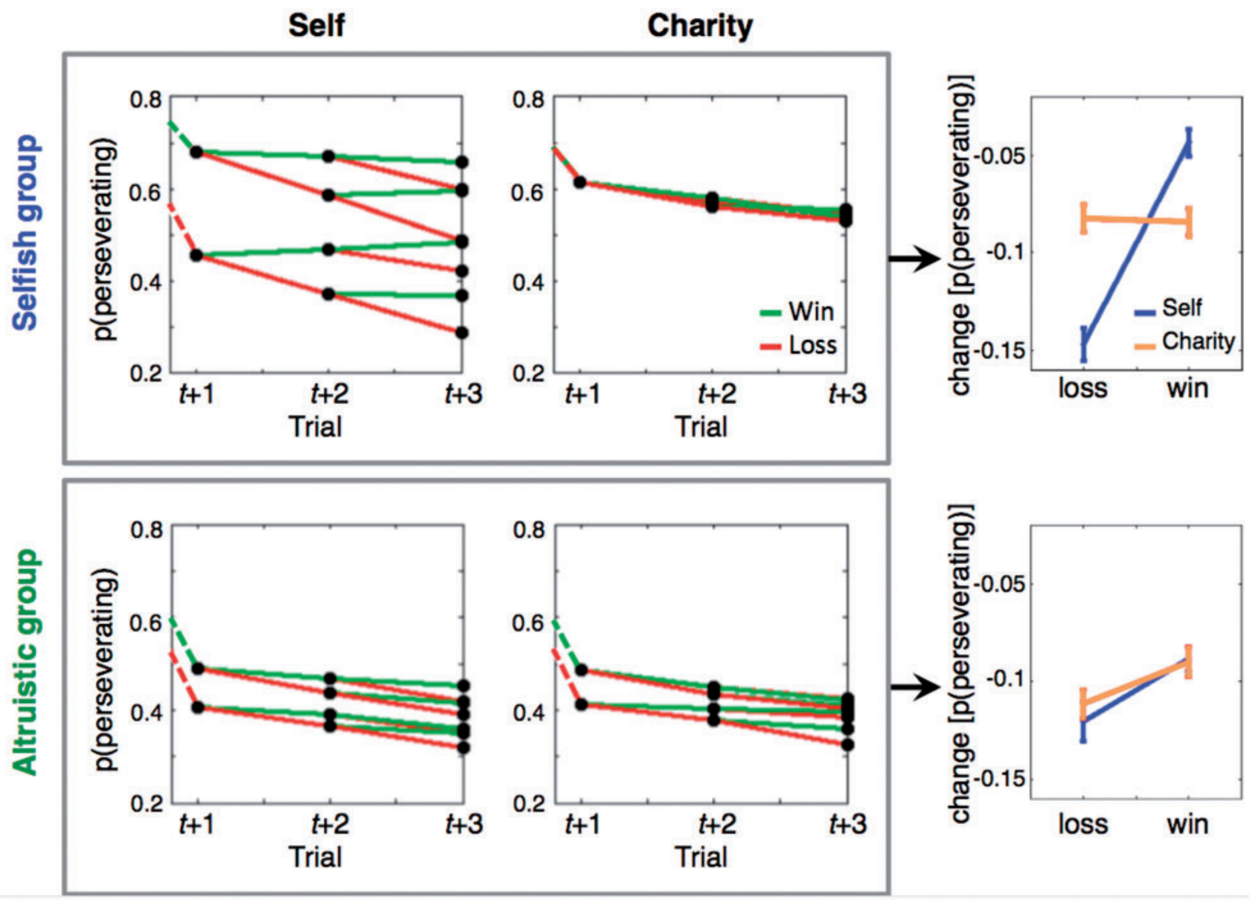

Fig. 3. Behavioral adjustment related to gains and losses for self and for the charity. For participants in the selfish-group the probability of persevering on a particular choice across trials was affected by outcomes for self but not outcomes for the charity. For participants in the altruistic-group, outcomes for self and outcomes for the charity had similar effects on behavioral adjustment. Error bars correspond to the standard error of the means (SE).

Downloaded from http://scan.oxfordjournals.org/ at Universidad Diego Portales on May 31, 2016

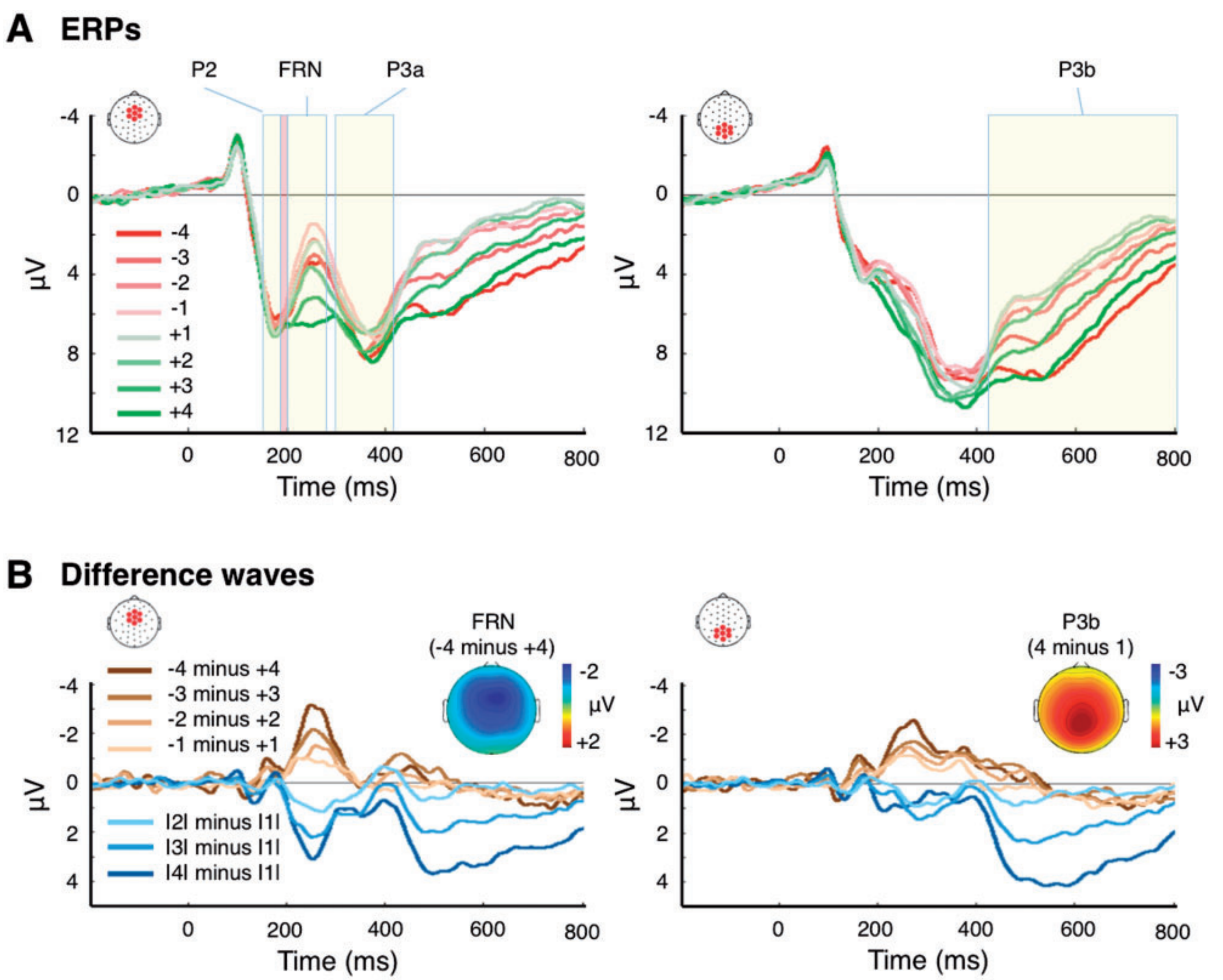

Fig 4. Effect of outcome valence and outcome magnitude on ERPs and difference waves. **(A)** Left, Grand average ERPs for each outcome from an anterior ROI of seven electrodes centered on FCz and located over frontocentral cortex. Cream-colored rectangles show the time windows defined for each of the ERP components. The thin rectangle in light red, immediately preceding the FRN time window, represents the anchoring reference latency window for quantifying the FRN amplitude. Right, Grand average ERPs from each outcome from a more posterior ROI of seven electrodes centered on Pz and located over the parietal cortex. **(B)** ERP difference waves obtained for the fronto-central ROI (left panel) and for the parietal ROI (right panel). Brown traces show the difference waves obtained by subtracting the ERP response elicited by a gain outcome from that elicited by a loss outcome, separately for each outcome magnitude. Blue traces show the difference waves obtained by subtracting the ERP response elicited by outcome magnitudes 2, 3 and 4 from that elicited by outcome magnitudes 1, regardless of outcome valence.

selfish-group adjusted their behavior based on gains and losses to themselves, regardless of the consequences for charity. For participants in the altruistic-group, self-outcomes and charity-outcomes seem to have a much less differentiated impact on behavioral adjustments. A repeated-measures ANOVA on the observed probability of persevering in a particular choice across trials (from trial t to trial $t+3$) corroborated these observations, reflected by a significant interaction between group, recipient and valence [F(1, 36)=18.9, $P<0.001$]. Specifically, for participants in the selfish-group, self-losses were associated with a larger decrease in the probability of perseverating than self-gains ($P<0.01$), but charity-losses and charity-gains did not have a significantly different impact on such probability ($P=0.86$). The ANOVA also showed a main effect of valence, because losses were associated with less perseveration than gains [F(1, 36)=65.5, $P<0.001$], regardless of group or recipient.

An alternative way of analysing this trial-by-trial behavioral-adjustment dynamic would be to use a shorter time horizon, by considering each trial t in terms of the chosen deck (S) and then calculating the observed probability of choosing S on trial $t+1$ (i.e. one sequential trial only, rather than three) as a function of whether the transitions between trials corresponded to gains or losses, separately for self and for the charity. As expected, and as observed in our supplementary material (Supplementary Figure S1), this method confirmed that participants in the selfish-group adjusted their behavior on the next trial based solely on outcomes for self, whereas participants in the altruistic-group adjusted their behavior based both on outcomes for self and outcomes for the charity.

Differential ERP responses to outcomes for self and outcomes for charity

Our ERP analyses examined whether participants in the altruistic-group differed from participants in the selfish-group in their neural responses to monetary outcomes for themselves *vs* for charity. We measured four ERP components (P2, FRN, P3a and P3b, Figure 4 for time windows and ROIs) and submitted each

Downloaded from http://scan.oxfordjournals.org/ at Universidad Diego Portales on May 31, 2016

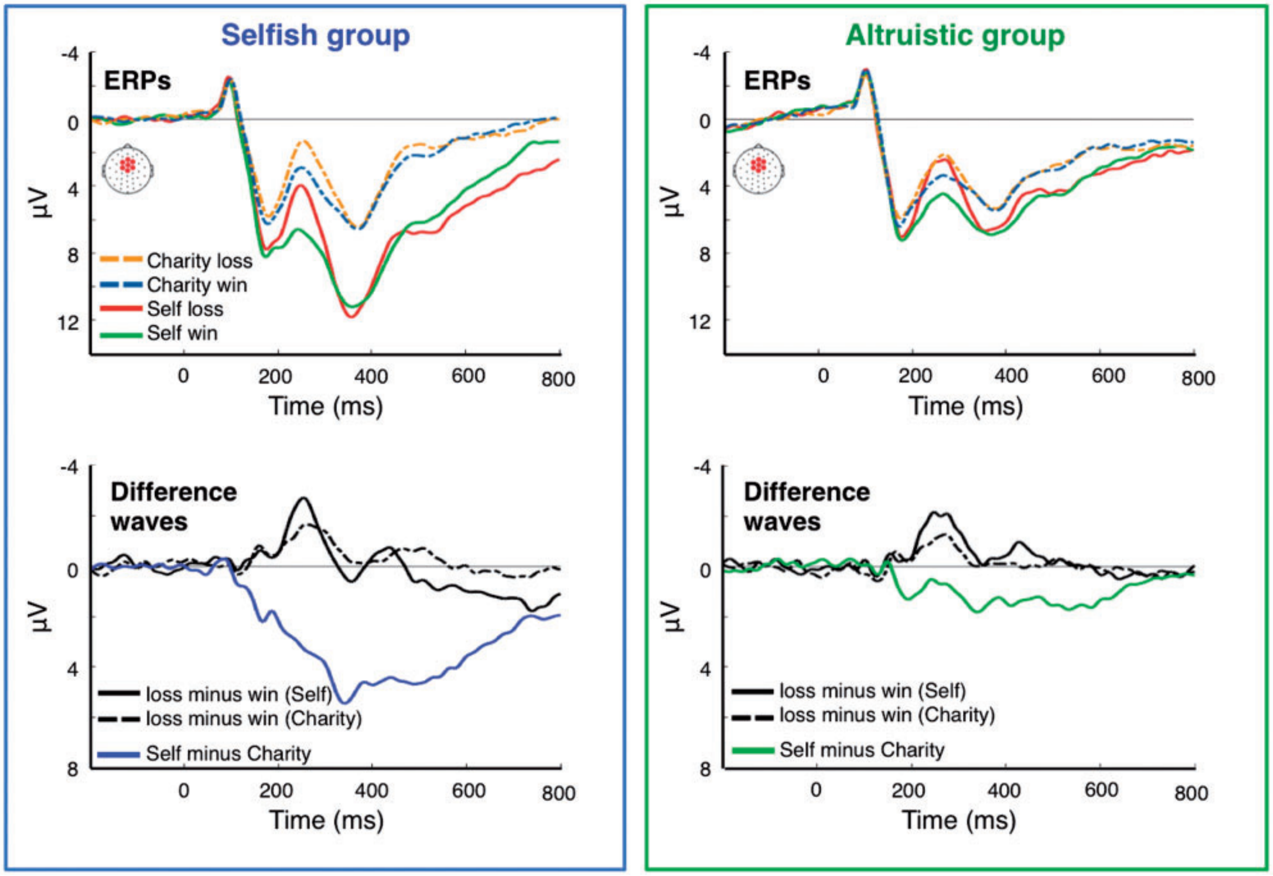

Fig. 5. ERP traces and loss-minus-win difference waves in the fronto-central ROI showing effects involving the factors 'group', 'recipient' and 'outcome valence', shown separately for the selfish group (**left panels**) and the altruistic group (**right panels**). **Upper panels:** Grand ERP averages for each outcome valence (gain outcome and loss outcomes) separately for each recipient (self and charity). **Lower panels:** Black traces show difference waves obtained by subtracting the ERP responses to gains from the ERP responses to losses, separately for each outcome recipient (self and charity). Blue and green traces show difference waves obtained by subtracting the ERP responses to outcomes for self from the ERP responses to outcomes for charity.

component to a repeated-measures ANOVA with four within-subject factors: valence (loss/win), magnitude (1–4), behavioral adjustment (switch/stay) and recipient (self/charity); and with group as a between-subjects factor (altruistic-group/selfish-group). Below we report the results of this analysis for each of these ERP components.

Effects on the P2 component (latency 152–184 ms). The ANOVA in our P2 measure did not find a significant effect of valence, magnitude or behavioral adjustment (P's> 0.05). It did find a main effect of outcome recipient [F(1, 36)=13.2, P< 0.001], indicating that outcomes for self-elicited a larger P2 response than outcomes for charity. Interestingly, although group did not have a main effect on the P2, the effect of recipient interacted with group [F(1, 36)=5.6, P< 0.05]. Subsequent *post hoc* contrasts revealed that this interaction derived from the P2 amplitude being greater in response to outcomes for self than for charity in participants within the selfish-group ($P < 0.01$), but with no significant self-*vs*-charity P2 difference in participants within the altruistic-group ($P = 0.83$) (Figure 5).

Given that within each trial the outcome for self could appear either before or after the outcome for charity, we examined whether the differences between groups on the neural responses to those outcomes might be explained by their relative sequence. Accordingly, we ran a complementary ANOVA on the outcome processing with recipient (self/charity) and sequence (presented first/presented second) as within-subject factors, and group as a between-subjects factor (altruistic-group/selfish-group). We found that both the main effect of recipient [F(1, 36)=15.6, $P < 0.001$] and the interaction between recipient and group [F(1, 36)=7.3, $P < 0.05$] continued to hold after controlling for the relative timing of outcomes for self and outcomes for charity within each trial. We also found a main effect of sequence [F(1, 36)=15.0, $P < 0.001$]. More specifically, the P2 was larger for the second outcome in the sequence than for the first one. However, we did not further analyse this effect because it was not related to our main experimental questions and it did not interact with factors recipient or group.

Effects on the FRN component (latency 204–272 ms). As expected, the FRN was modulated by outcome valence, being larger (i.e. more negative) after losses compared with gains [F(1, 36)=23.9, $P < 0.001$]. We also found a main effect of outcome magnitude [F(1, 108)=35.3, P< 0.001]. As confirmed by *post hoc* pairwise comparisons (1 > 2, 2 > 3, 3 > 4; P's < 0.05), the smaller the outcome magnitude, the larger the FRN (see left panel in Figure 4A). The ANOVA also revealed an interaction between outcome magnitude and outcome valence [F(1, 108)=5.7, $P < 0.05$]. In order to explore this interaction, we calculated the loss-minus-win FRN-difference for each magnitude (1 to 4, see left panel in Figure 4B)

Downloaded from http://scan.oxfordjournals.org/ at Universidad Diego Portales on May 31, 2016

on each participant, and submitted these measures to a one-way ANOVA, which revealed significant differences between these [F(3, 111)=5.7, $P < 0.05$]. *Post hoc* pairwise contrasts revealed that outcome magnitudes 4 and 3 were associated with larger loss-minus-win FRN-differences than outcome magnitudes 1 and 2 ($4 > 1$, $4 > 2$, $3 > 1$, $3 > 2$; P's < 0.01). No significant differences were found between outcome magnitudes 3 and 4 ($P = 0.66$), or between outcome magnitudes 1 and 2 ($P = 0.98$).

We did not find any significant effect of recipient, group or behavioral adjustment. However, we did find an interaction between outcome valence and recipient [F(1, 36)=7.0, $P < 0.05$]. We performed a *post hoc* analysis to explore this interaction. Specifically, we calculated, for each participant, the loss-minus-win FRN-difference separately for outcomes for self and outcomes for charity. The loss-minus-win FRN difference was significantly larger for self than for charity [t(74) = −2.52, $P < 0.05$] (see lower panels in Figure 5). Importantly, the original ANOVA did not find a three-way interaction between valence, recipient and group ($P = 0.2$), which indicates that the loss-minus-win FRN-difference was greater for self than for charity, regardless of whether participants were relatively more selfish or altruistic, and this FRN effect did not differ between the two groups.

Effects on the P3a component (latency 284–412 ms). The ANOVA performed on our frontally distributed P3a measure did not find a significant effect of valence or magnitude. However, the analysis revealed a main effect of behavioral adjustment [F(1, 36)=22.2, $P < 0.001$]. The P3a was significantly larger for outcomes that were followed by a different choice (i.e. switch trials) than for outcomes that were followed by the same choice (i.e. stay trials) (Supplementary Figure S2). This result is consistent with recent studies indicating that the P3 reflects processes that predict a future adjustment in choice behavior (Chase *et al.*, 2011; San Martín *et al.*, 2013; Zhang *et al.*, 2013).

The ANOVA performed on the P3a also revealed a main effect of recipient [F(1, 36)=57.4, $P < 0.001$], indicating that outcomes for self-elicited a larger P3a response than did outcomes for charity. The between-subjects factor of group was not associated to a significant main effect. However, similar to the results for the P2 and unlike the FRN, the effect of recipient interacted significantly with group [F(1, 36)=13.8, $P < 0.001$; Figure 5]. *Post hoc* contrasts revealed more specifically that the P3a amplitude was greater in response to outcomes for self than for charity in participants within the selfish-group ($P < 0.01$). In the altruistic group this trend approached significance ($P = 0.06$). Interestingly, the interaction between recipient and group was specifically explained by an increase of the P3a for self in the selfish-group compared with the P3a for self in the altruistic-group ($P < 0.05$), because there was no significant differences between these groups in their P3a responses to the outcomes for charity ($P = 0.96$) (Figure 5). Both the main effect of recipient [F(1, 36)=57.4, $P < 0.001$] and the interaction between recipient and group [F(1, 36)=14.6, $P < 0.001$] held after controlling for the relative within-trial presentation timing (first *vs* second) of the outcome for self *vs* for charity.

Effects on the P3b component (latency 416–796 ms). The ANOVA on the parietally distributed P3b revealed main effects of both outcome valence [F(1, 36)=6.4, $P = 0.0126$] and outcome magnitude [F(3, 108)=109.61, $P < 0.001$], indicating that the P3b was both greater for losses than gains and greater for large magnitude outcomes compared with small magnitude ones ($4 > 3$, $4 > 2$, $4 > 1$, $3 > 2$, $3 > 1$, P's < 0.001; but $2 > 1$, $P = 0.41$. See right panel in Figure 4B). We also found a main effect of behavioral adjustment [F(1, 36)=13.6, $P < 0.001$], indicating that the P3b was larger for switch trials than stay trials.

Similar to the results for the P2 and the P3a, we found a main effect of recipient on the P3b [F(1, 36)=70.4, $P < 0.001$], with larger responses being elicited by outcomes for self than by outcomes for charity (Supplementary Figure S3). Also similar to the results for the P2 and the P3a, group did not have a significant effect on the P3b ($P > 0.05$) and recipient interacted with group [F(1, 36)=16.2, $P < 0.001$]. However, and unlike the case of the P2 and P3a, the P3b was significantly larger for self than for charity's outcomes both in the selfish-group ($P < 0.01$) and the altruistic-group ($P < 0.05$). Specifically, while there were no significant differences between these groups with regard to their P3b responses to the outcomes for self ($P = 0.09$), the P3b in response to the outcomes for charity was significantly smaller in the selfish-group than in the altruistic-group ($P < 0.01$). The main effect of recipient [F(1, 36)=68.5, $P < 0.001$], and the interaction between recipient and group [F(1, 36)=16.7, $P < 0.001$], held after controlling for the relative timing within trials of the outcome for self and for charity.

For the P3b (and not for the P2 and P3a) we also found a three-way interaction between recipient, behavioral adjustment and group [F(1, 36)=4.8, $P < 0.05$]. Subsequent *post hoc* comparisons showed more specifically that only the P3b to outcomes for self in participants in the selfish-group significantly predicted behavioral adjustment (switch > stay after outcomes for self in the selfish-group P's < 0.01; $P > 0.4$ for the rest of the *post hoc* contrasts) (Supplementary Figure S2).

Finally, considering that P2, P3a and P3b correspond to different subcomponents of the P3 response, we conducted a supplementary repeated measures ANOVA with P3-subcomponent as a factor to evaluate whether the differences between these subcomponents would still hold when all were included in the same model. The results of that analysis matched the results that we obtained using a separate ANOVA for each subcomponent (Supplementary Table S1).

ERP responses to monetary feedback predict differences in self-reported helping behavior

The previous ERP analyses showed that the early stages of the P3 response, namely the P2 and P3a, distinguished between participants in the selfish-group and participants in the altruistic-group in that only selfish participants presented a significant self-*vs*-charity difference in these early subcomponents. This was not the case for the P3b, because even though the self-*vs*-charity P3b difference was smaller in the altruistic-group than in the selfish-group, even participants in the altruistic-group presented a significantly increased P3b response for self *vs* charity.

For these reasons, we specifically wanted to evaluate whether the P2 and P3a responses also predicted individual differences in actual donations (resulting from participants' performance) and self-reports of engagement in real-life altruistic behaviors. For each participant, we extracted the difference between the mean amplitude of the P2 and P3a responses to outcomes for charity *vs* for outcomes to self (i.e. self-minus-charity ERP amplitude differences) and evaluated the participant-wise correlation between these ERP contrasts and both actual-donations and the altruism subscore of the HOQ. Across participants, the correlations between the P2 amplitude and the altruism subscore ($r_{(36)} = -0.32$, $P = 0.05$) and actual-donations ($r_{(36)} = -0.30$, $P = 0.07$) approached significance. The P3a contrast significantly covaried with both the altruism subscore ($r_{(36)} = -0.42$, $P = 0.004$) and the actual-donations ($r_{(36)} = -0.43$, $P = 0.003$) (Figure 6). In summary, the greater

Downloaded from http://scan.oxfordjournals.org/ at Universidad Diego Portales on May 31, 2016

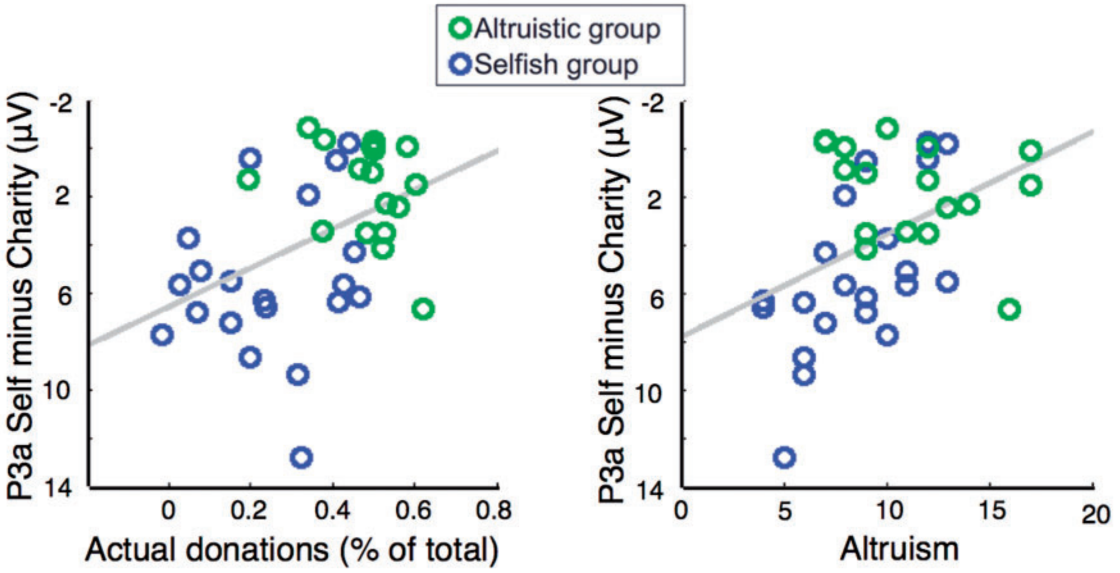

Fig. 6. The across-participants association between the self-minus-charity-P3a contrasts and both actual-donations (**left**) and self-reported real-life altruism (**right**). Participants included in the selfish group for previous analyses are represented as blue rings. Participants included in the altruistic group for previous analyses are represented as green rings.

Table 1. Summary of effects across ERP components

	P2	FRN	P3a	P3b
ANOVA (F, *p*)				
Recipient (self > charity)	13.2, <0.001	0.61, 0.44	57.4, <0.001	70.4, <0.001
Recipient × group (self > charity, greater in the selfish group)	5.55, 0.02	0.81, 0.37	13.8, <0.001	16.2, <0.001
Valence × recipient (loss > win, greater for self)	0.78, 0.38	7.02, 0.01	3.40, 0.07	3.84, 0.06
Valence × recipient × group (loss > win, greater for self in the selfish group)	0.20, 0.66	1.73, 0.20	0.10, 0.75	2.34, 0.14
Correlations (r, *p*)				
Actual donations	−0.3, 0.07	−0.19, 0.25	−0.43, 0.004	. . .
Reported real life altruism	−0.32, 0.05	−0,20, 0.22	−0.42, 0.003	. . .

Notes F- Values are in Roman, P-Values are in Italics.

the self-minus-charity-P3a amplitude across participants, the smaller the actual-donations resulting from the task and the lower the self-reported probability of engaging real-life altruism behaviors, with the same general tendencies being observed for the earlier P2. In contrast, performing the same analysis with the FRN feedback valence effect (loss minus win for self *vs* loss minus win for charity) did not show a significant association with either altruism or actual-donations (P's > 0.22). We summarize this and our other key results in Table 1.

Discussion

In social environments, decisions determine rewards not only for oneself but also for others. Recent behavioral findings have indicated that individual differences in the ability to learn from outcomes for others (i.e. social learning) can explain part of the variation in altruism observed among people (Kwak *et al.*, 2014). The present work indicates specific neurocognitive mechanisms—as indexed by ERP components—that may underlie the link between altruism and social learning. We showed that the amplitude of the P3 and its subcomponents, which have been associated with attention-related processes related to learning and environmental model updating, predicted individual differences both in social learning and in self-reported engagement in real-life altruistic behaviors. Specifically, the less altruistic the participants, the greater the difference in the P3 response to outcomes for self *vs* outcomes for the charity. In contrast, no relationships between altruistic tendencies and the FRN component were observed.

Our findings suggest that individual differences in pro-social behavior are not associated with differences in the way that the brain extracts utility from outcomes for oneself *vs* for others (as reflected by the FRN), but with differences in the amount of attention that is allocated to learn from outcomes for oneself *vs* outcomes for others (P3). Moreover, the fact that altruism-related differences in the P3 response are observable as early as the P2 time window (∼170 ms) suggest that the differences in altruistic tendencies are associated with early processes of attentional capture. Overall, the present electrophysiological results provide insights into how individual differences in pro-social behavior are associated with differences in the way that the brain allocates attentional resources for the processing of outcomes for oneself *vs* for others.

Altruism is associated with differential reward learning for self and others

Previous behavioral studies have shown that there are significant individual differences in the expression of pro-social

Downloaded from http://scan.oxfordjournals.org/ at Universidad Diego Portales on May 31, 2016

behaviors (Fehr and Schmidt, 2003; Murphy *et al.*, 2011) and in the neurocognitive mechanisms that have been linked to that heterogeneity (Tankersley *et al.*, 2007; Donaldson and Young, 2008; Morishima *et al.*, 2012). In a recent behavioral study, Kwak *et al.* (2014) showed that this heterogeneity is also associated with individual differences in the learning of pro-social reward contingencies. They found that altruistic individuals, compared with more selfish ones, were better at learning to increase monetary gains for a charity. Our behavioral results replicate and extend this finding by showing that the intended-donations declared by participants before a reward-learning task covaried both with their actual-donations resulting from performance in the task and with their self-reported engagement in real-life altruistic behavior as measured by the HOQ. Moreover, our behavioral results also show that participants with more selfish goals adjusted their decision behavior based on gains and losses for themselves, with little regard to the consequences for charity. In contrast, participants with more altruistic goals adjusted their behavior based on the consequences for both themselves and for charity. Overall, our behavioral results suggest that participants' goals were aligned with their daily life altruistic tendencies, and that those goals strongly influenced their final results in the task.

Unexpectedly, participants with more altruistic goals were also worse at avoiding the deck that had a negative EV for both self and charity (S−\C−). One potential explanation for this result is that attempting to learn about both self and charity outcomes—as was clearly intended by these participants—increases the overall difficulty of the task compared with tracking only one's own outcomes. Critically, this hypothesis predicts that participants who are altruistic to the point of disregarding outcomes for self should be just as good at avoiding the S−\C− choice as selfish participants who disregard outcomes for the charity. Unfortunately, we could not test this possibility here because few participants' behavior approached that extreme sort of altruism (Figure 2).

The FRN distinguishes other-regarding outcomes but not altruistic tendencies

Previous studies have shown that different social contexts can modulate subjects' empathic responses toward others and that the FRN ERP component covaries with such empathic responses (Yu and Zhou, 2006; Itagaki and Katayama, 2008; Fukushima and Hiraki, 2009). From this, we expected to find FRN responses (reflecting losses minus wins) both to outcomes for oneself and to outcomes for charity. We also may have expected that the difference between the magnitudes of these effects (FRN to outcomes for self minus FRN to outcomes for the charity) would covary with participants' self-reported altruistic tendencies, with this difference being smaller for those with higher self-reported levels of such tendencies. The results showed a difference between the loss-minus-win FRN responses for self *vs* charity outcomes, although the relative sizes of these effects were unaffected by altruism. Below, we discuss the implications of these results.

Crucially, losses (*vs* wins) for the charity elicited an FRN in both groups of participants, indicating that neural calculations of monetary outcomes affecting societal causes share processing similarities to outcomes during interpersonal interactions (Yu and Zhou, 2006; Itagaki and Katayama, 2008; Fukushima and Hiraki, 2009; Leng and Zhou, 2010; Ma *et al.*, 2011). Given that previous neuroimaging studies have found that transfers to charity elicit neural activity in subcortical areas that typically respond to primary rewards for the individual (Moll *et al.*, 2006; Harbaugh *et al.*, 2007; Hare *et al.*, 2010), our results also provide indirect support to the theory that the FRN reflects cortical activity that is being modulated by utility computations performed by subcortical reward areas. This suggestion should be taken with caution, however, because these subcortical circuits, in contrast to cortical regions, do not have the appropriate geometry to be able produce ERPs at the scalp (Luck, 2005), and thus the ERP recordings do not directly reflect activity in subcortical circuits. On the other hand, due to the high temporal resolution of the ERPs employed here, the current results extend previous neuroimaging findings by indicating that the human brain calculates the value from contributing to both self-related and societal causes at a relatively early stage of processing—namely, within ~250 ms of the presentation of that outcome.

It is also the case that the FRN feedback valence effects were larger for self *vs* charity, for both the selfish and altruistic groups. Moreover, these self-*vs*-charity-ERP effects did not differ between the two groups and did not predict individual differences in self-reported altruism, intended donations, nor actual donations derived from the task. This suggests that individual differences in pro-social behavior are not associated with differences in the way that the brain extracts utility from outcomes for oneself *vs* for others. This negative result may seem surprising given that a previous study found that an antagonist's gain *vs* loss elicited a negative-going FRN, as if receiving such information was being interpreted as a loss of some sort to oneself (Itagaki and Katayama, 2008). From this we may have expected to see a tendency toward a reversed polarity FRN in response to charity's outcomes, especially in our group of selfish participants. However, the FRN feedback valence effect was greater for self than charity, and similarly so regardless of participants' altruism.

Our results thus suggest that the FRN reflects an initial and relatively coarse evaluation of the acquired/lost value that may distinguish between 'self' and 'others', but which does not covary with participants' altruistic tendencies. This interpretation may seem to conflict with findings indicating that the FRN toward others' outcomes is modulated by empathy (Yu and Zhou, 2006; Fukushima and Hiraki, 2009). However, previous studies have also suggested that brain responses underlying altruistic tendencies could actually be independent from brain responses underlying empathy (Tankersley *et al.*, 2007). Although we did not include any psychometric empathy scales in our study, our results are consistent with this latter suggestion.

The P3 predicts individual differences in altruism

In contrast to the FRN, we found large differential amplitudes of the P3 responses for self *vs* charity that covaried robustly with individual differences in intended donations, actual donations and self-reported engagement in real life altruistic behaviors. Importantly all the P3 subcomponents (i.e. P2, P3a, P3b) were larger in response to outcomes for self than to outcomes for the charity in both the selfish and altruistic groups. However, and specially with regard to the early stages of the P3 response (i.e. P2, P3a), that difference was even greater for the selfish individuals than for the more altruistic ones. This finding indicates that participants in the selfish and altruistic groups differ in the allocation of attentional resources to outcomes for oneself *vs* outcomes for others.

The fact that altruism-related differences in attention allocation are observable as early as the P2 deflection (~170 ms) suggest that the differences in altruistic tendencies are associated with an attentional process that precedes the evaluation of

Downloaded from http://scan.oxfordjournals.org/ at Universidad Diego Portales on May 31, 2016

feedback valence (~240 ms for the FRN). The P2 has been shown to be specifically associated with higher arousal levels (Carretié *et al.*, 2001; Schutter *et al.*, 2004) and attention capturing by target stimuli (Potts *et al.*, 1996, 2006; Potts, 2004). This raises the possibility that low levels of altruism may be associated with the presence of a feature-specific bottom-up saliency map like those that appear to underlie some attentional capture effects during visual search (Eimer *et al.*, 2009; Zhang *et al.*, 2012). Salience coding could selectively enhance the signal-to noise ratio of preferred outcomes (e.g. those generating advantages rewards for oneself). Testing such a conjecture could be an important direction for future work.

In addition, the P3b in the selfish group was larger in response to outcomes for self that preceded changes in choice behavior than for outcomes that were followed by staying with the same choice in the next trial. This result suggests that participants in the selfish groups attend only the outcomes for oneself during the cognitive updating of represented cue-reward contingencies. Strengthening this interpretation, this ERP result presents a striking parallel with our behavioral results showing that 'selfish' participants adjusted their behavior according to outcomes for self with no regard to the outcomes for charity.

Taken together, our P3 results are broadly consistent with the view of the P3 as an ERP component reflecting the amount of cognitive resources used during the revision of an internal model of the environment that is used to achieve subjective goals (Donchin, 1981; Donchin and Coles, 1988). This 'context-updating theory' hypothesis for the P3 predicts that the amplitude of such ERP activity will scale with the probability of adjusting choices on subsequent trials, a result that has been reported by previous studies (Chase *et al.*, 2011; San Martín *et al.*, 2013; Zhang *et al.*, 2013) and was also observed here. It indeed seems likely that the participants in our study dynamically updated an internal model of the association between the deck symbols and the probability of winning for self and for the charity. Under this view, decreasing levels of altruism across participants would be associated with an increasing amount of attention devoted to outcomes for themselves and learning from those outcomes, as compared with attention toward outcomes for the charity. This would then be reflected in a negative correlation between the self-minus-charity P3 contrast and the donations to the charity resulting from participants' choices during the task, as was the case. This interpretation is also consistent with results from the previous behavioral study by Kwak *et al.* (2014), who found that a relatively higher weighting of charitable as opposed to personal outcomes, as reflected in the reinforcement learning model parameter, predicted altruistic behavior.

In our study, intended donations, actual performance-based donation, self-reported altruism and differential P3 responses for self *vs* charity appear to be intimately related to each other. Although we cannot trace the full set of links among these variables, we can propose two causal interpretations that could be tested in future studies: First, daily life altruistic tendencies, as measured by self-reports, may have led to setting a subjective goal in our task, as measured by intended donations, which may in turn have determined the relative amounts of attention paid to outcome for self and outcome for charity, which would then be reflected in the differential self *vs* charity P3. Consistent with this interpretation, Moll *et al.* (2006) found that altruism was associated with the activation of anterior regions of the medial prefrontal cortex that have also been implicated in goal representation (Wood and Grafman, 2003; Kringelbach and Radcliffe, 2005).

A second explanation would attribute a causal role to attention-related learning biases by saying that innate or acquired individual differences in the ability to learn about outcomes for others, as measured by the differential self *vs* charity P3, would act as a cognitive precursor for real-life altruistic tendencies. These tendencies may, in turn, determine the relative amounts of attention devoted to learn how to increase the benefits for charity *vs* self. In this sense, the ability to learn about rewards for others could be part of the neural substrate of altruism. Indirect support for this interpretation comes from the fact that a likely contributing neural source of the P3 is the temporal-parietal junction (TPJ) (Kiss *et al.*, 1989; Smith *et al.*, 1990; Halgren *et al.*, 1995), a region which acts as a hub of streams that support cognitive processes such as attention, memory and social processing (Carter and Huettel, 2013). Indeed, TPJ activity has been shown to be involved in shifting attention to focus on another's perspective (Saxe and Kanwisher, 2003; Behrens *et al.*, 2008; Hampton *et al.*, 2008; Young and Saxe, 2009). In addition, activity in this region correlates with subjects' self-reported altruism (Tankersley *et al.*, 2007) and willingness-to-give as measured by average donations (Hare *et al.*, 2010). Moreover, individual differences in gray matter volume in the TPJ have been shown to scale with individuals' altruism (Morishima *et al.*, 2012).

Importantly, the two above proposed interpretations are not necessarily incompatible with each other. It could indeed be the case that innate or acquired biases in learning about rewards for others constitute a precursor for altruism, which would lead to altruistic goals that shape learning about benefits for others. Tracing the interaction and causal relationships between these factors will be an important pursuit for future studies aimed at deciphering the nature and neural underpinnings of human altruism.

Acknowledgements

We thank Janaina Brizante, Eshita Singh, and Charles Woldorff for helpful assistance with data collection.

Funding

This work was supported by a Fulbright/CONICYT scholarship to R.S.M., [NIMH R01-86712] to S.A.H., and [NIMH R01-NS051048] to M.G.W.

Supplementary data

Supplementary data are available at *SCAN* online.

Conflict of interest. None declared.

References

Apicella, P., Legallet, E., Trouche, E. (1997). Responses of tonically discharging neurons in the monkey striatum to primary rewards delivered during different behavioral states. *Experimental Brain Research*, **116**, 456–66.

Batson, C.D., Shaw, L.L. (1991). Evidence of prosocial motives toward a pluralism for altruism. *Psychological Inquiry*, **2**, 107–22.

Behrens, T.E.J., Hunt, L.T., Woolrich, M.W., Rushworth, M.F.S. (2008). Associative learning of social value. *Nature*, **456**, 245–9.

Bell, A.J., Sejnowski, T.J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, **7**, 1129–59.

Downloaded from http://scan.oxfordjournals.org/ at Universidad Diego Portales on May 31, 2016

Billeke, P., Zamorano, F., Cosmelli, D., Aboitiz, F. (2012). Oscillatory brain activity correlates with risk perception and predicts social decisions. *Cerebral Cortex*, **23**, 2872–83.

Carretié, L., Mercado, F., Tapia, M., Hinojosa, J. (2001). Emotion, attention, and the "negativity bias", studied through event-related potentials. *International Journal of Psychophysiology*, **41**, 75–85.

Carter, R.M., Huettel, S. (2013). A nexus model of the temporal-parietal junction. *Trends in Cognitive Sciences*, **17**, 328–36.

Chase, H.W., Swainson, R., Durham, L., Benham, L., Cools, R. (2011). Feedback-related negativity codes prediction error but not behavioral adjustment during probabilistic reversal learning. *Journal of Cognitive Neuroscience*, **23**, 936–46.

Debener, S., Ullsperger, M., Siegel, M., Fiehler, K., von Cramon, D.Y., Engel, A.K. (2005). Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. *Journal of Neuroscience*, **25**, 11730–7.

Delgado, M.R. (2007). Reward-related responses in the human striatum. *Annals of New York Academy of Sciences*, **1104**, 70–88.

Delgado, M.R., Nystrom, L.E., Fissell, C., *et al.* 2014. Tracking the hemodynamic responses to reward and punishment in the striatum feedback that confirms reward expectation triggers auditory cortex activity tracking the hemodynamic responses to reward and punishment in the striatum. *Journal of Neurophysiology*, **84,** 3072–7.

Delorme, A., Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, **134**, 9–21.

Delorme, A., Sejnowski, T., Makeig, S. (2007). Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *Neuroimage*, **34**, 1443–9.

Donaldson, Z.R., Young, L.J. (2008). Oxytocin, vasopressin, and the neurogenetics of sociality. *Science*, **322**, 900–4.

Donchin, E. (1981). Surprise!.Surprise? *Psychophysiology. Psychophysiology*, **18**, 493–513.

Donchin, E., Coles, M.G.H. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences*, **11**, 357–74.

Eichele, T., Specht, K., Moosmann, M., *et al.* (2005). Assessing the spatiotemporal evolution of neuronal activation with single-trial event-related potentials and functional MRI. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 17798–803.

Eimer, M., Kiss, M., Press, C., Sauter, D. (2009). The roles of feature-specific task set and bottom-up salience in attentional capture: an ERP study. *Journal of Experimental Psychology. Human Perception and Performance*, **35**, 1316–28.

Fehr, E., Fischbacher, U. (2003). The nature of human altruism. *Nature*, **425**, 785–91.

Fehr, E., Schmidt, K. 2003. Theories of fairness and reciprocity: evidence and economic applications. In: Dewatripont M, Hansen LP, Turnovsky SJ, editors. *Advances in Economics and Econometrics*, 208–57. Cambridge: Cambridge University Press.

Fukushima, H., Hiraki, K. (2009). Whose loss is it? Human electrophysiological correlates of non-self reward processing. *Society for Neuroscience*, **4**, 261–75.

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

Halgren, E., Baudena, P., Clarke, J.M., *et al.* (1995). Intracerebral potentials to rare target and distractor auditory and visual stimuli. II. Medial, lateral and posterior temporal lobe. *Electroencephalography and Clinical Neurophysiology*, **94**, 229–50.

Hamilton, W.D. (1964). The genetical evolution of social behaviour. II. *Journal of Theoretical Biology*, **7**, 17–52.

Hampton, A.N., Bossaerts, P., O'Doherty, J.P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 6741–6.

Harbaugh, W.T., Mayr, U., Burghart, D.R. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science*, **316**, 1622–5.

Hare, T., Camerer, C.F., Knoepfle, D.T., Rangel, A. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *Journal of Neuroscience*, **30**, 583–90.

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**, 679–709.

Itagaki, S., Katayama, J. (2008). Self-relevant criteria determine the evaluation of outcomes induced by others. *Neuroreport*, **19**, 383–7.

Jung, T.P., Makeig, S., Humphries, C., *et al.* (2000a). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, **37**, 163–78.

Jung, T.P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E., Sejnowski, T.J. (2000b). Removal of eye activity artifacts from visual event-related potentials in normal and clinical subjects. *Clinical Neurophysiology*, **111**, 1745–58.

Kiss, I., Dashieff, R.M., Lordeon, P. (1989). A parieto-occipital generator for P300: evidence from human intracranial recordings. *International Journal of Neuroscience*, **49**, 133–9.

Knutson, B., Peterson, R. (2005). Neurally reconstructing expected utility. *Games and Economic Behavior*, **52**, 305–15.

Kringelbach, M.L., Radcliffe, J. 2005. The human orbitofrontal cortex: linking reward to hedonic experience, 6, 691–702.

Kwak, Y., Pearson, J.M., Huettel, S.A. (2014). Differential reward learning for self and others predicts pro-social behavior. *PLoS One*, **9**, 1–9.

Leng, Y., Zhou, X. (2010). Modulation of the brain activity in outcome evaluation by interpersonal relationship: an ERP study. *Neuropsychologia*, **48**, 448–55.

Luck, S. 2005. *An Introduction to the Event-Related Potential Technique*. Cambridge, MA: MIT Press.

Ma, Q., Shen, Q., Xu, Q., Li, D., Shu, L., Weber, B. (2011). Empathic responses to others' gains and losses: an electrophysiological investigation. *Neuroimage*, **54**, 2472–80.

Mayr, U., Harbaugh, W.T., Tankersley, D. 2008. Neuroeconomics of charitable giving and philanthropy. In: Glimcher PW, Camerer CF, Fehr E, Poldrack RA, editors. *Neuroeconomics: Decision Making and the Brain*, 303–20. New York, USA: Academic Press.

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

Moll, J., Krueger, F., Zahn, R., Pardini, M., de Oliveira-Souza, R., Grafman, J. (2006). Human fronto-mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 15623–8.

Morishima, Y., Schunk, D., Bruhin, A., Ruff, C.C., Fehr, E. (2012). Linking brain structure and activation in temporoparietal junction to explain the neurobiology of human altruism. *Neuron*, **75**, 73–9.

Murphy, R.O., Ackermann, K.A., Handgraaf, M.J.J. (2011). Measuring social value orientation. *Judgment and Decision Making*, **6**, 771–81.

Downloaded from http://scan.oxfordjournals.org/ at Universidad Diego Portales on May 31, 2016

Nieuwenhuis, S. 2011. Learning, the P3, and the locus coeruleus-norepinephrine system. In: Mars RB, Sallet J, Rushworth MFS, Yeung N, editors. *Neural Basis of Motivation and Cognitive Control*, 209–22. Cambridge, MA: MIT Press.

Nieuwenhuis, S., Aston-Jones, G., Cohen, J.D. (2005). Decision making, the P3, and the locus coeruleus-norepinephrine system. *Psychological Bulletin*, **131**, 510–32.

Nieuwenhuis, S., Yeung, N., Holroyd, C.B., Schurger, A., Cohen, J.D. (2004). Sensitivity of electrophysiological activity from medial frontal cortex to utilitarian and performance feedback. *Cerebral Cortex*, **14**, 741–7.

Nowak, M.A., Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, **437**, 1291–8.

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clinical Neurophysiology*, **118**, 2128–48.

Potts, G.F. (2004). An ERP index of task relevance evaluation of visual stimuli. *Brain and Cognition*, **56**, 5–13.

Potts, G.F., Liotti, M., Tucker, D.M., Posner, M.I. (1996). Frontal and inferior temporal cortical activity in visual target detection: evidence from high spatially sampled event-related potentials. *Brain Topography*, **9**, 3–14.

Potts, G.F., Martin, L.E., Burton, P., Montague, P.R. (2006). When things are better or worse than expected: the medial frontal cortex and the allocation of processing resources. *Journal of Cognitive Neuroscience*, **18**, 1112–9.

Rangel, A., Camerer, C., Montague, P.R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience*, **9**, 545–56.

Rigoni, D., Polezzi, D., Rumiati, R., Guarino, R., Sartori, G. (2010). When people matter more than money: an ERPs study. *Brain Research Bulletin*, **81**, 445–52.

Romer, D., Cruder, C.L., Lizzadro, T. (1986). A person-situation approach to altruistic behavior. *Journal of Personality and Social Psychology*, **51**, 1001–12.

San Martín, R., Appelbaum, L.G., Huettel, S.A., Woldorff, M.G. (2016). Cortical brain activity reflecting attentional biasing toward reward-predicting cues covaries with economic decision-making performance. *Cerebral Cortex*, **26**, 1–11.

San Martín, R. (2012). Event-related potential studies of outcome processing and feedback-guided learning. *Frontiers in Human Neuroscience*, **6**, 1–17.

San Martín, R., Appelbaum, L.G., Pearson, J.M., Huettel, S.A., Woldorff, M.G. (2013). Rapid brain responses independently predict gain maximization and loss minimization during economic decision making. *Journal of Neuroscience*, **33**, 7011–9.

San Martín, R., Manes, F., Hurtado, E., Isla, P., Ibañez, A. (2010). Size and probability of rewards modulate the feedback error-related negativity associated with wins but not losses in a monetarily rewarded gambling task. *Neuroimage*, **51**, 1194–204.

Saxe, R., Kanwisher, N. (2003). People thinking about thinking people the role of the temporo-parietal junction in "theory of mind." *Neuroimage*, **19**, 1835–42.

Scheibe, C., Ullsperger, M., Sommer, W., Heekeren, H.R. (2010). Effects of parametrical and trial-to-trial variation in prior probability processing revealed by simultaneous electroencephalogram/functional magnetic resonance imaging. *Journal of Neuroscience*, **30**, 16709–17.

Schutter, D.J.L.G., de Haan, E.H.F., van Honk, J. (2004). Functionally dissociated aspects in anterior and posterior electrocortical processing of facial threat. *International Journal of Psychophysiology*, **53**, 29–36.

Smith, M.E., Halgren, E., Sokolik, M., *et al.* (1990). The intracranial topography of the P3 event-related potential elicited during auditory oddball. *Electroencephalography and Clinical Neurophysiology*, **76**, 235–48.

Tankersley, D., Stowe, C.J., Huettel, S. (2007). Altruism is associated with an increased neural response to agency. *Nature Neuroscience*, **10**, 150–1.

van Schie, H.T., Mars, R.B., Coles, M.G.H., Bekkering, H. (2004). Modulation of activity in medial frontal and motor cortices during error observation. *Nature Neuroscience*, **7**, 549–54.

Woldorff, M.G., Liotti, M., Seabolt, M., Busse, L., Lancaster, J.L., Fox, P.T. (2002). The temporal dynamics of the effects in occipital cortex of visual-spatial selective attention. *Brain Research Cognitive Brain Research*, **15**, 1–15.

Wood, J.N., Grafman, J. (2003). Human prefrontal cortex: processing and representational perspectives. *Nature Reviews Neuroscience*, **4**, 139–47.

Yeung, N., Sanfey, A.G. (2004). Independent coding of reward magnitude and valence in the human brain. *Journal of Neuroscience*, **24**, 6258–64.

Young, L., Saxe, R. (2009). An FMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience*, **21**, 1396–405.

Yu, R., Zhou, X. (2006). Brain responses to outcomes of one's own and other's performance in a gambling task. *Neuroreport*, **17**, 1747–51.

Yu, R., Zhou, X. (2009). To bet or not to bet? The error negativity or error-related negativity associated with risk-taking choices. *Journal of Cognitive Neuroscience*, **21**, 684–96.

Zhang, D., Gu, R., Wu, T., *et al.* (2013). An electrophysiological index of changes in risk decision-making strategies. *Neuropsychologia*, **51**, 1397–407.

Zhang, X., Zhaoping, L., Zhou, T., Fang, F. (2012). Neural activities in v1 create a bottom-up saliency map. *Neuron*, **73**, 183–92.

Downloaded from http://scan.oxfordjournals.org/ at Universidad Diego Portales on May 31, 2016