Social deviance activates the brain's error-monitoring system

Bo-Rin Kim · Alison Liss · Monica Rao ·
Zachary Singer · Rebecca J. Compton

Published online: 25 October 2011
© Psychonomic Society, Inc. 2011

Abstract Social psychologists have long noted the tendency for human behavior to conform to social group norms. This study examined whether feedback indicating that participants had deviated from group norms would elicit a neural signal previously shown to be elicited by errors and monetary losses. While electroencephalograms were recorded, participants ($N = 30$) rated the attractiveness of 120 faces and received feedback giving the purported average rating made by a group of peers. The feedback was manipulated so that group ratings either were the same as a participant's rating or deviated by 1, 2, or 3 points. Feedback indicating deviance from the group norm elicited a feedback-related negativity, a brainwave signal known to be elicited by objective performance errors and losses. The results imply that the brain treats deviance from social norms as an error.

Keywords ERP · Motivation · Cognitive control

The present study was designed to address the neurocognitive mechanisms contributing to social deviance detection and conforming behavior. Social conformity, the tendency of behavior to change in accordance with social group norms, has long been a favored topic of social psychologists (e.g., Asch, 1951; Festinger, 1954). Conformity can be considered a type of goal-directed action, in which the goals of behavior include maximizing

the rewards that follow from accurate performance and social acceptance (Aronson, 2004; Cialdini & Goldstein, 2004). We hypothesized that detecting social deviance and engaging in corrective action depend on the same neurocognitive mechanisms that are known to be involved in goal-directed actions in nonsocial domains. Recent studies have fleshed out a cognitive control system, involving the anterior cingulate and dorsolateral prefrontal cortex, that monitors ongoing performance, compares performance outcomes with expected outcomes, and generates warning signals in response to errors or conflicts in performance (e.g., Botvinick, 2007; Holroyd & Coles, 2002; Ridderinkhof, Ullsperger, Crone, & Nieuwenhuis, 2004). These error signals are thought to be incorporated into adaptive plans for corrective action.

We hypothesized that social deviance would elicit an error signal from the same brain system that monitors for errors in other performance domains. Researchers have defined a brainwave potential, the *error-related negativity* (ERN), that is measurable from electroencephalogram (EEG) recordings within 100 ms following the commission of performance errors such as an erroneous buttonpress (Falkenstein, Hohnsbein, Hoormann, & Blanke, 1991; Gehring, Goss, Coles, Meyer, & Donchin, 1993). A similar potential, called the *feedback-related negativity* (FRN), occurs approximately 200–400 ms after explicit negative feedback indicating errors or monetary losses in tasks for which the accuracy of the response is not self-evident (e.g., Luu, Tucker, Derryberry, Reed, & Poulsen, 2003; Miltner, Braun, & Coles, 1997). Both the ERN and the FRN are thought to be generated by the anterior cingulate cortex (Mathalon, Whitfield, & Ford, 2003; Miltner et al., 1997; Swick & Turken, 2002; Ullsperger & von Cramon, 2006). Furthermore, the amplitudes of the ERN and FRN predict

B.-R. Kim · A. Liss · M. Rao · Z. Singer · R. J. Compton (✉)
Department of Psychology, Haverford College,
370 Lancaster Avenue,
Haverford, Pennsylvania 19041, USA
e-mail: rcompton@haverford.edu

subsequent corrective behavior, implying that the error signal feeds into a process of adaptive behavior change (e.g., Cohen & Ranganath, 2007; Debener et al., 2005; Gehring et al., 1993).

We predicted that the FRN signal would be elicited by feedback indicating that an individual's judgment deviated from the group norm. Such evidence would suggest that the neural control system known to be operative in cognitive performance tasks is more broadly applicable to contexts of social cognition as well. Preliminary evidence supporting this prediction comes from an fMRI study that found increased anterior cingulate cortex activity when participants received feedback indicating social deviance (Klucharev, Hytönen, Rijpkema, Smidts, & Fernández, 2009). We sought to extend these prior findings with evidence from EEG methods. Whereas anterior cingulate activity, as measured by fMRI, may reflect a range of mental processes, including arousal, attentional control, and action selection (e.g., Paus, 2001; Rushworth, Buckley, Behrens, Walton, & Bannerman, 2007), the FRN is more specifically associated with error monitoring.

Method

Participants

Participants ($N = 30$ undergraduates; 18 female, 12 male) had normal or corrected-to-normal vision, no neurological history, and no current use of psychoactive medications.

Attractiveness rating task

Participants rated the attractiveness of female faces in a task modeled after that of Klucharev et al. (2009). To create the stimulus set, 200 color photographs from a high school

yearbook were rated for attractiveness and distinctiveness in an untimed rating task completed by 19 pilot participants. From this set, the 120 faces with the most average attractiveness ratings were selected ($M = 4.86$ on a 9-point scale), after excluding faces with high levels of distinctiveness (ratings >6 on a 9-point scale).

The task instructions indicated that after rating each face, the participant would see the average rating given by students in the same college community. Participants were informed that their ratings would be used to determine which faces would be included in a database for the psychology departments of the Quaker Consortium (a consortium of local colleges) and were instructed to respond as accurately as possible.

Figure 1 presents the sequence of events for each trial. First, a face appeared for 250 ms, followed by a prompt for the participant's rating. The rating was made on a 1 (*least attractive*) to 9 (*most attractive*) scale using the number pad on a computer keyboard. This rating then appeared for 500 ms below the word "Yours." Then the participant's rating appeared for 1 s together with the purported group rating, which was positioned below the word "Average" (feedback screen). Finally, the two ratings appeared together with the face for 2 s, to reinforce the relationship between the rating discrepancy and the specific face.

The task included 120 trials, with feedback rigged on each trial. Of the trials, 30 showed group ratings that differed from the participant's rating by ± 3 points, 30 showed group ratings that differed by ± 2 points, 30 showed group ratings that differed by ± 1 point, and 30 trials showed group ratings that were the same as the participant's rating (0 discrepancy). Specific faces were randomly paired with feedback levels during the construction of the experiment, and these pairings were constant across participants. Trials (face–feedback pairings) were presented in an intermixed order that was randomized separately for each participant.

Fig. 1 Sequence of trial events, depicted schematically. Actual face photographs were presented rather than the smiley-face icon. A 500-ms fixation screen preceded each trial

The number of trials contributing to each plus or minus feedback condition varied somewhat across participants, due to a procedure used to correct for off-scale feedback responses. There were always 30 trials per discrepancy magnitude (e.g., 30 trials for ± 3 -point feedback). Of these 30 trials, 15 were initially programmed to include plus feedback (a group rating higher than the participant's), and 15 were programmed to include minus feedback (a group rating lower than the participant's). However, if the feedback was computed to result in an off-scale value—for example, +3 feedback on a trial in which the participant gave a rating of 7—the sign of the feedback was reversed (e.g., in this case, the group rating would be given as 4 rather than 10). Therefore, depending on the actual responses of the participant, there was not always an even distribution of plus and minus trials for a given feedback discrepancy level. This was particularly the case for 3-pt discrepancy trials, because these were the most likely to need correction for off-scale feedback. Nevertheless, across subjects, the average number of trials for each of the discrepancy and feedback direction (plus or minus) conditions was 15 (see Table 1). Prior research had found that the ERN was relatively stable and reliable with as few as 6–8 trials (Olvet & Hajcak, 2009; Pontifex et al., 2010; see also Marco-Pallares, Cucurell, Münte, Strien, & Rodriguez-Fornells, 2011).

Ratings were presented in color to highlight the discrepancy between the participant's rating, which was always in red, and the group rating. Group ratings that were the same as the participant's rating were red, ratings that differed by ± 1 were red-orange, ratings that differed by ± 2 were yellow-orange, and ratings that differed by ± 3 were yellow.

Following the main rating task and EEG cap removal, participants rated the same 120 faces again. In this second run, faces were presented in an untimed fashion and in random order, and no feedback was given. The difference in ratings between the first and second runs served as the measure of conformity.

Table 1 Numbers of trials contributing to feedback conditions separated by direction and magnitude of feedback discrepancy

Feedback Discrepancy	Mean	SD	Range
+1	15.2	1.10	13–19
–1	14.8	1.10	11–17
+2	14.9	2.60	9–22
–2	15.1	2.60	8–21
+3	14.5	3.98	10–25
–3	15.5	3.98	5–20

EEG recording and signal processing

Data were recorded continuously from four midline scalp sites (Fz, FCz, Cz, Pz) using sintered Ag/AgCl electrodes embedded in a Quik-Cap system. These four scalp sites were chosen because they are the sites at which the FRN is typically observed (see, e.g., Gehring & Willoughby, 2002; Miltner et al., 1997). Signals were amplified by a NuAmps amplifier controlled by Neuroscan software, with a sampling rate of 1000 Hz and a bandpass of 0.1–40 Hz. Data were referenced online to the right mastoid and digitally rereferenced offline to the average of the left and right mastoids. Eye movements were monitored by electrodes above and below the left eye and at the outer canthus of each eye. Upon visual inspection, portions of the EEG record with large nonblink artifacts were manually excluded. The effect of blinks was reduced using the Neuroscan software's regression-based algorithm for ocular artifact reduction. Finally, epochs were rejected if scalp voltages exceeded a $\pm 150\text{-}\mu\text{V}$ threshold.

Epochs surrounding feedback onset were defined by the interval from -200 to 600 ms, with Time 0 as the onset of the feedback screen. Amplitude values between -200 and 0 ms were used for baseline correction. For each participant, epochs were averaged separately for each feedback condition (discrepancies of 0, ± 1 , ± 2 , and ± 3 rating points) to create the event-related waveforms.

Results

Attractiveness ratings

Figure 2 presents the mean scores for the change in attractiveness ratings as a function of feedback. Overall,

Fig. 2 Change in ratings (second rating minus first rating) as a function of feedback discrepancy. A feedback discrepancy of -3 indicates that the feedback screen gave the group rating as 3 points lower than the participant's own rating, whereas a feedback discrepancy of 3 indicates that the group rating presented was 3 points higher than the participant's own rating. Error bars show ± 1 SEM

change scores tended to be negative [$M = -0.70$, $SEM = 0.09$; one-sample t test, $t(29) = -7.55$, $p < .001$], indicating lower ratings of attractiveness in the second relative to the first run.

More importantly, the change in ratings depended on the feedback discrepancy, as confirmed by a one-way repeated measures ANOVA on change scores, with feedback discrepancy as the factor [$F(6, 174) = 31.8$, $p < .0001$]. The linear contrast was significant [$F(1, 29) = 90.1$, $p < .0001$], indicating a linear relationship between feedback discrepancy and changes in ratings. Participants' ratings in the second run decreased the most when the purported group rating was 3 points below the participant's original rating, and decreased the least when the group rating was 3 points above the participant's original rating. These results imply behavioral conformity with the purported group rating.

Individual differences in conformity, as quantified by the slope relating feedback discrepancy to rating change (mean slope = 0.22), were unrelated to individual differences in the mean decrease in rating from Time 1 to Time 2 ($r = -.14$, $p > .45$). That is, the pattern of conformity behavior (the slope of the function) was independent from the overall decrease in ratings (the intercept of the function). The general decrease in ratings from Times 1 to 2 may have resulted from fatigue, repetition effects, or different viewing conditions at Time 1 (brief exposure) than at

Time 2 (untimed exposure), but this overall decrease was not relevant to the main focus of the study. Individual differences in conformity (ratings slope) were also unrelated to the proportions of trials in the plus versus minus feedback discrepancy conditions ($r = .15$, $p > .40$), confirming that the number of trials per condition was an irrelevant nuisance variable.

FRN data

Effect of feedback discrepancy Figure 3 depicts grand-average waveforms for the four feedback conditions and four electrode sites. Waveforms are time-locked to feedback onset, rather than to response onset, because the response-locked ERN could not logically be evoked in a paradigm in which participants cannot evaluate the "correctness" of an action prior to receiving feedback. The figure displays a feedback-related divergence in the waveforms that begins around 200 ms postfeedback, lasts until approximately 450 ms, and is superimposed upon a large positive-going waveform (cf. Miltner et al., 1997). To quantify this effect statistically, mean amplitudes in the 200- to 450-ms window were submitted to a 4 (feedback condition: discrepancy of 0, ± 1 , ± 2 , ± 3) $\times$ 4 (electrode site: Fz, FCz, Cz, Pz) repeated measures ANOVA. The Greenhouse–Geisser

Fig. 3 Grand-average waveforms at four electrode sites for feedback conditions in which the group average rating differed from the participant's rating by 0, 1, 2, or 3 points. Time 0 is the onset of the feedback screen

correction was applied to p values in this and all subsequent analyses to correct for violations of sphericity.

Supporting the hypothesis, the main effect of feedback was significant [$F(3, 87) = 37.6, p < .0001$]. The means (see Fig. 4) indicated less positive amplitudes for the discrepant-feedback conditions than for the nondiscrepant condition (Fisher's least significant difference [LSD] post-hoc test, $ps < .001$). Among the discrepant conditions, there was some gradation of response, such that amplitudes were less positive for the ± 2 and ± 3 conditions than for the ± 1 condition (LSD test, $ps < .05$), although the ± 2 and ± 3 conditions did not differ significantly from one another.

The response to feedback differed according to the site [main effect of site, $F(3, 87) = 8.4, p < .01$; site $\times$ feedback, $F(9, 261) = 4.7, p < .001$; see Table 2]. At each site, the simple main effect of feedback was significant [$F(3, 87) > 20, ps < .0001$], and the 0-discrepancy condition differed from all three discrepant conditions (LSD post-hoc, $ps < .05$). However, some sites distinguished the degree of feedback discrepancy better than others did. At the Fz site, the means for the three discrepant conditions ($\pm 1, \pm 2, \pm 3$) did not differ from one another. In contrast, at the FCz, Cz, and Pz sites, the amplitudes were less positive for the ± 2 and ± 3 conditions than for the ± 1 condition (LSD post-hoc, $ps < .05$). These site-related effects are consistent with evidence that sensitivity to the magnitude of feedback information has a somewhat more posterior distribution than the FRN itself (Goyer, Woldorff, & Huettel, 2008).

Because the apparent FRN is superimposed upon a large positivity—presumably the P300 component, on the basis of its timing, shape, and scalp distribution—we conducted an additional analysis on trough-to-peak FRN amplitudes to confirm that the feedback effect was not attributable to variations in the P300. In this analysis, we computed the FRN amplitude by subtracting the most negative value in the 250- to 350-ms window (FRN peak) from the most positive value in the 150- to 250-ms window (preceding

Fig. 4 Mean amplitude in the 200- to 450-ms window, averaged across sites, as a function of feedback discrepancy. Error bars indicate one standard error of the mean

Table 2 Mean (and *SEM*) feedback-related negativity amplitudes (in microvolts), as a function of feedback discrepancy and electrode site

Feedback Discrepancy	Electrode Site			
	Fz	FCz	Cz	Pz
0	7.30 (0.78)	8.82 (0.90)	8.28 (0.99)	6.14 (0.97)
± 1	4.57 (0.76)	5.39 (0.80)	4.92 (0.83)	3.72 (0.96)
± 2	3.93 (0.76)	4.40 (0.77)	3.57 (0.82)	1.94 (0.85)
± 3	4.21 (0.75)	4.17 (0.71)	2.98 (0.71)	1.74 (0.71)
Average	5.00 (0.70)	5.70 (0.73)	4.94 (0.77)	3.39 (0.81)

trough). More-positive values on this index indicate greater FRN amplitudes. Consistent with the prior analysis and with the waveforms depicted in Fig. 3, this analysis also yielded a statistically significant effect of feedback [$F(3, 87) = 11.12, p < .001$]. FRN amplitudes were 3.53, 4.95, 5.32, and 6.08 μV for feedback discrepancies of 0, 1, 2, and 3 points, respectively, constituting a significant linear trend [$F(1, 29) = 23.51, p < .001$]. Post-hoc tests on the main effect means confirmed that each discrepant-feedback condition differed from the zero-discrepancy condition (LSD, $ps < .005$), the ± 1 condition differed from the ± 3 condition ($p < .05$), and the ± 2 condition did not differ from the ± 1 condition but did differ marginally from the ± 3 condition ($p < .08$). These trough-to-peak findings confirmed the main pattern that emerged in the analysis of mean amplitudes.

Effect of feedback direction A second analysis examined the effect of the direction of feedback discrepancy on the FRN response. This analysis separated trials not only by the amount (absolute value) of discrepancy between the participant's rating and the purported group rating, as in the prior analysis, but also by the direction of the discrepancy—that is, whether the group rating was higher or lower than the participant's rating. The ANOVA on mean amplitudes included the within-subjects factors of discrepancy magnitude (1, 2, or 3 points), discrepancy direction (plus or minus; *plus* indicates a group rating higher than the participant's rating), and site (Fz, FCz, Cz, and Pz). Zero-discrepancy trials were not included in this analysis because they could not be categorized by direction.

The ANOVA revealed a significant main effect of discrepancy direction [$F(1, 29) = 5.34, p < .03$]. Amplitudes were more negative for plus feedback conditions ($M = 3.36 \mu V, SEM = 0.73$), in which the group rating was higher than the participant's rating, as compared to minus feedback conditions ($M = 4.18 \mu V, SEM = 0.65$), in which the group rating was lower than the participant's rating. This effect is illustrated in the waveforms in Fig. 5. Thus, the FRN response was sensitive to the direction of the discrepancy

Fig. 5 Grand-average waveforms at the FCz site for feedback conditions in which the group rating was higher than the participant's rating (plus condition) or lower than the participant's rating (minus condition). No-difference (zero-discrepancy) trials are shown for purposes of comparison

between the individual and group ratings, and not only to its magnitude.

The ANOVA also revealed significant main effects of discrepancy magnitude [$F(2, 58) = 6.45, p < .005$] and site [$F(3, 87) = 7.48, p < .005$], as well as a site $\times$ discrepancy magnitude interaction [$F(6, 174) = 3.80, p < .02$]. These results have already been described in the previous analysis.

Between-subjects prediction of conformity behavior The next set of analyses sought to determine whether individual differences in conformity behavior were associated with individual differences in the neural response to social feedback. To quantify conformity behavior for each participant, we calculated for each of them the slope of the function relating feedback discrepancy and the change in rating (see Fig. 2). A steeper positive slope indicates greater influence of the feedback on the subsequent ratings (that is, more conformity), whereas a shallower slope indicates less influence of the feedback (i.e., less conformity). The mean slope was 0.22, significantly greater than 0 [$t(29) = 9.49, p < .001$], as would be expected based on the group data in Fig. 2. Across participants, slopes ranged from -0.03 to 0.52 .

The ratings slope was added as a continuous predictor to the ANOVA on mean FRN amplitudes, with feedback condition (0, ± 1 , ± 2 , ± 3) and site (Fz, FCz, Cz, and Pz) as factors. One participant was excluded due to FRN amplitude values more than 3 SDs higher than the group mean. Inclusion of this participant in the analyses of individual differences led to spurious significant results that disappeared when the participant was excluded.

The analyses revealed no significant effects involving the ratings slope predictor. Most importantly, the interaction between feedback condition and ratings slope was not significant ($F < 1$). Thus, there was no evidence that the neural response to discrepant feedback differed systematically among participants whose conformity behavior showed either steeper or shallower slopes.

Within-subjects prediction of ratings change The goal of this analysis was to determine whether there was a within-subjects relationship between the amplitude of the FRN and the subsequent change in ratings. Trials with 2-pt and 3-pt feedback discrepancies were divided in two parts, separately for each participant, on the basis of whether the faces on those trials subsequently received a greater change in attractiveness ratings in the direction of the feedback (more-change trials) or a lesser change in rating in the direction of the feedback (less-change trials). The division of trials into “more-change” and “less-change” categories was carried out separately for the 2-pt and 3-pt discrepancy conditions, because otherwise 3-pt discrepancy trials would tend to load onto the more-change category and 2-pt discrepancy trials onto the less-change category, confounding the purpose of the analysis. Trials with zero discrepancy were not included in this analysis, nor were trials with 1-pt discrepancies, because these trials produced both smaller FRNs and less behavioral change, and therefore would be expected to be less sensitive to feedback effects.

Mean amplitudes were extracted for more-change and less-change trials and were submitted to an ANOVA with trial type (more change, less change) and site as repeated measures factors. The main effect of trial type was not significant ($F < 1$), nor was the interaction effect ($F < 1$). Therefore, FRN amplitudes were neither significantly higher nor lower on trials that later produced more behavioral change.

Discussion

The main results of this study demonstrate that the brain responds to deviance from social group norms similarly to how it responds to feedback indicating incorrect performance or monetary losses, even when there is no objectively correct answer on the task. Feedback indicating that the participants' ratings of facial attractiveness were discrepant from group norms produced a significant divergence in the event-related waveforms, as compared to feedback indicating consistency between the participants' ratings and group norms. The shape, polarity, and timing of this divergence were highly similar to those of the FRN, a brainwave component known to be elicited by errors and losses (e.g., Gehring & Willoughby, 2002; Miltner et al., 1997). This component occurred within a few hundred milliseconds after the onset of feedback, indicating that the brain's error-monitoring system responds rapidly to evidence of social deviance. The FRN was sensitive to the amount of discrepancy from group norms, exhibiting a stronger negativity for 2- or 3-point discrepancies than for discrepancies of only 1 point, and it was

sensitive to the direction of discrepancy from group norms, exhibiting a stronger negativity when the purported group rating fell above rather than below the participant's rating.

The present results extend prior findings that demonstrated a similar pattern using fMRI (Klucharev et al., 2009). Conceptual replication by different methodologies, aside from further substantiating the robustness of the phenomenon, can take advantage of the different strengths of different methodologies. fMRI findings can provide stronger evidence of the spatial localization of social effects in the anterior cingulate cortex; for example, one fMRI study found that social and nonsocial feedback produced activation in slightly different regions of the cingulate cortex (Behrens, Hunt, Woolrich, & Rushworth, 2008). Conversely, the ERP method carries its own advantages. First, the present data provide information about the timing of the neural response to deviance, demonstrating that it occurs quickly, beginning within approximately 200 ms after the onset of the feedback information. Furthermore, whereas fMRI activity in the anterior cingulate cortex has been associated with a wide range of cognitive states (e.g., Paus, 2001; Rushworth et al., 2007), the present FRN finding links the neural response to deviance to more specific cognitive operations involved in behavioral control. According to the theory posited by Holroyd and Coles (2002), the FRN reflects a cognitive process that determines when outcomes are worse than expected, indicating an absence of expected reward, and that generates an error signal that is used in reinforcement learning. Because social acceptance is inherently rewarding and social exclusion is punishing (e.g., Baumeister & Leary, 1995; Blanton & Christie, 2003), the FRN in response to social deviance may mark the potential loss of social reward and contribute to social learning through reinforcement. Alternatively, since some prior research has found that the FRN can be elicited by both positive and negative outcomes that are unexpected (Oliveira, McDonald, & Goodman, 2007), participants may simply expect to be typical, and the FRN in response to deviance may mark a violation of that expectancy.

Although the most plausible interpretation of the present findings is that participants categorize their own social deviance as an error, alternative explanations need to be considered. For example, both the ERN and the FRN can be evoked by observing errors committed (or losses sustained) by other people (e.g., van Schie, Mars, Coles, & Bekkering, 2004; Yu & Zhou, 2006). Therefore, it is possible that participants view the discrepancy between their own rating and the group rating as an error made by the group, rather than as an error made by the participant. Although this possibility cannot be completely ruled out, it seems inconsistent with the finding that participants' ratings of the faces changed in a manner indicating conformity

with the group ratings, a pattern that is difficult to explain if participants believed that the group ratings were incorrect. Nevertheless, this alternative explanation could be pursued in future studies. In addition, although the task was a subjective rating task with no objectively correct or incorrect answer, it was possible that participants felt pressure to obtain group consensus because the experimenters stressed the importance of the data, purportedly for a common database to be used in future studies. This explanation does not undermine the main conclusion that the FRN is sensitive to social feedback, but it raises questions for future research about the circumstances under which such feedback may be more or less compelling.

The findings also revealed that participants were sensitive to the direction of the feedback, generating larger FRN responses when the purported group consensus rated the face as more attractive than had the participant, as compared to when the group consensus rated the face as less attractive than had the participant. This adds further evidence of the sensitivity of the neural response to nuances of the feedback information. It also rules out less interesting alternative explanations for the main findings, such as the possibility that participants were only responding to the color discrepancy rather than to the meaning of the comparison between the participant's and group's rating. If the FRN had been evoked only in response to the conflicting colors of the participant's and the group's ratings, the same response would be expected for plus and minus feedback conditions, which were matched for color discrepancy. Therefore, the FRN difference between the plus- and minus-feedback conditions confirms that participants were attending to the meaning conveyed by the feedback information.

We had not specifically predicted a larger FRN to plus-feedback conditions as compared to minus-feedback conditions, so we can only speculate about the best interpretation of this effect. One possibility is that rating a face as less attractive (relative to the group rating) is deemed less socially acceptable than rating a face as more attractive (relative to the group rating). That is, participants might feel that they have been "too cruel" in judging a face as less attractive than the group. Another possibility is that participants may confuse the value of the rating with the value of their own performance, therefore viewing a lower rating relative to the group as somehow reflecting worse ("below average") performance on the part of the participant. Both of these possible interpretations imply that although technically there is no right or wrong answer on the rating task, participants may have implicit notions about which rating discrepancies are more acceptable. Prior studies have demonstrated that socially unacceptable race-biased errors produced larger ERNs than socially acceptable errors in an objective performance task

(Amodio et al., 2004). Therefore, it would not be surprising if concerns about social acceptability also influenced the neural response to feedback about social deviance in the attractiveness rating task.

Attempts to link FRN responses to conformity behavior produced mixed results. On the one hand, both the FRN and subsequent behavioral ratings responded significantly to the same experimental manipulation (the feedback discrepancy), implying that they are associated. Yet attempts to find correlations between the FRN amplitude and subsequent behavior change did not yield significant results. That is, participants who were more conforming, on the basis of the ratings slope, did not have either larger or smaller FRN responses overall in response to group feedback indicating deviance. Likewise, trials that produced more behavior change within a given participant (and within a discrepancy level) did not necessarily have larger or smaller FRN responses than did trials that produced less behavior change. Therefore, the data do not support a simple claim that a larger FRN may lead to more (or less) behavior change.

Indeed, even in the cognitive control literature, the relationship between the amplitude of an error signal and the subsequent behavioral change, such as error correction or posterror slowing, has been difficult to establish reliably (for a review, see Gehring, Liu, Orr, & Carp, 2012). Dissociations between error detection and behavior change may be consistent with two-step models of cognitive control. In such models, the cognitive control system includes an evaluative component that detects errors and conflicts, and a subsequent corrective component that brings about behavior change. These steps are thought to be instantiated in anterior cingulate cortex and dorsolateral prefrontal cortex (DLPFC), respectively (e.g., Botvinick, 2007; Ridderinkhof et al., 2004). A two-step model provides the opportunity for intervening variables to determine how exactly behavior is changed once errors or conflicts are detected, rather than assuming that detection of an error automatically leads to a certain behavior change. The lack of significant correlations between FRN and conformity behavior in the present study highlights the need to better understand how social deviance detection, in combination with other factors, may contribute to behavior change.

Future research can further examine whether the factors that influence social conformity affect the FRN elicited by social deviance or the brain's subsequent response to the FRN signal. A number of variables are known to affect the likelihood of conformity, including task importance, task difficulty, centrality of the social group to the participant's identity, and personality (e.g., Aronson, 2004; Baron, Vandello, & Brunsman, 1996; Cialdini & Goldstein, 2004; Rhodes & Wood, 1992). Some of these factors may influence the degree to which deviance is categorized as an "error," thereby presumably also affecting the amplitude of the FRN

response. For example, stronger FRN responses to social deviance may be elicited when the task is deemed more important, just as the ERN response to objective errors is more pronounced when accuracy is stressed (Gehring et al., 1993).

Other factors may impact the engagement of corrective action once social deviance is detected. For example, people who value nonconformity may detect deviance with equal FRN responses, but may not engage in behavioral conformity as a result of that deviance detection. Such individual-difference factors may affect DLPFC activity rather than the FRN, according to current models of cognitive control. Preliminary evidence has indeed suggested that activity in the DLPFC is associated with norm compliance (Spitzer, Fischbacher, Herrnberger, Grön, & Fehr, 2007). Thus, research using a neurocognitive approach could potentially provide a new way of categorizing variables that influence social conformity, on the basis of the stage of processing at which they have their largest impact.

Generally speaking, the results from the present study imply that the psychological processes that respond to social deviance may rely, at least in part, on the same neurocognitive systems that implement other aspects of cognitive control, such as trial-and-error learning and the adjustment of cognitive strategies following errors and conflict (Botvinick, 2007; Holroyd & Coles, 2002). The results emphasize that the brain has a general-purpose control system that can be applied to monitoring outcomes of behavior across many domains. Cognitive neuroscience research has already demonstrated ERN and FRN responses to errors and losses in many contexts—errors committed with hands and feet (Holroyd, Dien, & Coles, 1998), errors in buttonpressing and errors in speech (e.g., Masaki, Tanaka, Takasawa, & Yamazaki, 2001), or errors of commission and errors of timing (e.g., Luu, Flaisch, & Tucker, 2000; Miltner et al., 1997). The anterior cingulate cortex is also activated by cognitive dissonance (van Veen, Krug, Schooler, & Carter, 2009), implying that conflict between internal attitudes and behavior can produce an error-like neural response. The present results indicate that social "errors" also produce a neural response from the cingulate cortex, broadening the scope of this general-purpose control system even further.

Author note The study was conducted as a bachelor's thesis at Haverford College by the first four authors, who contributed equally and are listed alphabetically. We gratefully acknowledge comments from Benjamin Le, Joshua Carp, and Daniel Arnstein.

References

- Amodio, D. M., Harmon-Jones, E., Devine, P. G., Curtin, J. J., Hartley, S. L., & Covert, A. E. (2004). Neural signals for the detection of unintentional race bias. *Psychological Science*, 15, 88–93.

- Aronson, E. (2004). *The social animal*. New York: Worth.
- Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (Ed.), *Groups, leadership and men: Research in human relations* (pp. 177–190). Pittsburgh: Carnegie Press.
- Baron, R. S., Vandello, J. A., & Brunsman, B. (1996). The forgotten variable in conformity research: Impact of task importance on social influence. *Journal of Personality and Social Psychology*, 71, 915–927.
- Baumeister, R. F., & Leary, M. R. (1995). The need to belong: The desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117, 497–529.
- Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, 456, 245–249.
- Blanton, H., & Christie, C. (2003). Deviance regulation theory: A theory of action and identity. *Review of General Psychology*, 7, 155–149.
- Botvinick, M. M. (2007). Conflict monitoring and decision making: Reconciling two perspectives on anterior cingulate function. *Cognitive, Affective, & Behavioral Neuroscience*, 7, 356–366.
- Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, 55, 591–621.
- Cohen, M. X., & Ranganath, C. (2007). Reinforcement learning signals predict future decisions. *Journal of Neuroscience*, 27, 371–378.
- 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, 11730–11737.
- Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke, L. (1991). Effects of cross-modal divided attention on late ERP components: II. Error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology*, 78, 447–455.
- Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7, 117–140. doi:10.1177/001872675400700202
- Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, 4, 385–390.
- Gehring, W. J., Liu, Y., Orr, J. M., & Carp, J. (2012). The error-related negativity (ERN/Ne). In S. J. Luck & E. S. Kappenman (Eds.), *Oxford handbook of event-related potential components*. New York: Oxford University Press.
- Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, 295, 2279–2282.
- Goyer, J. P., Woldorff, M. G., & Huettel, S. A. (2008). Rapid electrophysiological brain responses are influenced by both valence and magnitude of monetary reward. *Journal of Cognitive Neuroscience*, 20, 2058–2069.
- 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. doi:10.1037/0033-295X.109.4.679
- Holroyd, C. B., Dien, J., & Coles, M. G. H. (1998). Error-related scalp potentials elicited by hand and foot movements: Evidence for an output-independent error processing system in humans. *Neuroscience Letters*, 242, 65–68.
- Klucharev, V., Hytönen, K., Rijpkema, M., Smidts, A., & Fernández, G. (2009). Reinforcement learning signal predicts social conformity. *Neuron*, 61, 140–151. doi:10.1016/j.neuron.2008.11.027
- Luu, P., Flaisch, T., & Tucker, D. M. (2000). Medial frontal cortex in action monitoring. *Journal of Neuroscience*, 20, 464–469.
- Luu, P., Tucker, D. M., Derryberry, D., Reed, M., & Poulsen, C. (2003). Electrophysiological responses to errors and feedback in the process of action regulation. *Psychological Science*, 14, 47–53.
- Marco-Pallares, J., Cucurell, D., Münte, T. F., Strien, N., & Rodríguez-Fornells, A. (2011). On the number of trials needed for a stable feedback-related negativity. *Psychophysiology*, 48, 852–860. doi:10.1111/j.1469-8986.2010.01152.x
- Masaki, H., Tanaka, H., Takasawa, N., & Yamazaki, K. (2001). Error-related brain potentials elicited by vocal errors. *Neuroreport*, 12, 1851–1855.
- Mathalon, D. H., Whitfield, S. L., & Ford, J. M. (2003). Anatomy of an error: ERP and fMRI. *Biological Psychology*, 64, 119–141.
- Miltner, W. H. R., Braun, C. H., & Coles, M. G. H. (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, 788–798.
- Oliveira, F. T. P., McDonald, J. J., & Goodman, D. (2007). Performance monitoring in the anterior cingulate is not all error-related: Expectancy deviation and the representation of action–outcome associations. *Journal of Cognitive Neuroscience*, 19, 1994–2004.
- Olvet, D. M., & Hajcak, G. (2009). The stability of error-related brain activity with increasing trials. *Psychophysiology*, 46, 957–961.
- Paus, T. (2001). Primate anterior cingulate cortex: Where motor control, drive and cognition interface. *Nature Reviews Neuroscience*, 2, 417–424.
- Pontifex, M. B., Scudder, M. R., Brown, M. L., O’Leary, K. C., Wu, C.-T., Themanson, J. R., Hillman, C. H. (2010). On the number of trials necessary for stabilization of error-related brain activity across the lifespan. *Psychophysiology*, 47, 767–773.
- Rhodes, N., & Wood, W. (1992). Self-esteem and intelligence affect influenceability: The mediating role of message reception. *Psychological Bulletin*, 111, 156–171. doi:10.1037/0033-2909.111.1.156
- Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science*, 306, 443–447. doi:10.1126/science.1100301
- Rushworth, M., Buckley, M., Behrens, T., Walton, M., & Bannerman, D. (2007). Functional organization of the medial frontal cortex. *Current Opinion in Neurobiology*, 17, 220–227.
- Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G., & Fehr, E. (2007). The neural signature of social norm compliance. *Neuron*, 56, 185–196.
- Swick, D., & Turken, A. U. (2002). Dissociation between conflict detection and error monitoring in the human anterior cingulate cortex. *Proceedings of the National Academy of Sciences*, 99, 16354–16359.
- Ullsperger, M., & von Cramon, D. Y. (2006). The role of intact frontostriatal circuits in error processing. *Journal of Cognitive Neuroscience*, 18, 651–664.
- 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, 549–554.
- van Veen, V., Krug, M. K., Schooler, J. W., & Carter, C. S. (2009). Neural activity predicts attitude change in cognitive dissonance. *Nature Neuroscience*, 12, 1469–1474.
- Yu, R., & Zhou, X. (2006). Brain responses to outcomes of one’s own and other’s performance in a gambling task. *Neuroreport*, 17, 1747–1751.