Introduction

Guilt is an experience that arises when we violate norms or values that we consider important—for example, when we have

final sample (total of $N = 43$) had normal or corrected-to-normal

Kédia et al. 2008; Cui et al.2015
; Lepron et al. 2015). Mistakes in the dot-estimation task

Fig. 1A), participants underwent two func-

Fig. 1B), participants played a similar dot-
and sadness, for each of the 6 experimental conditions.

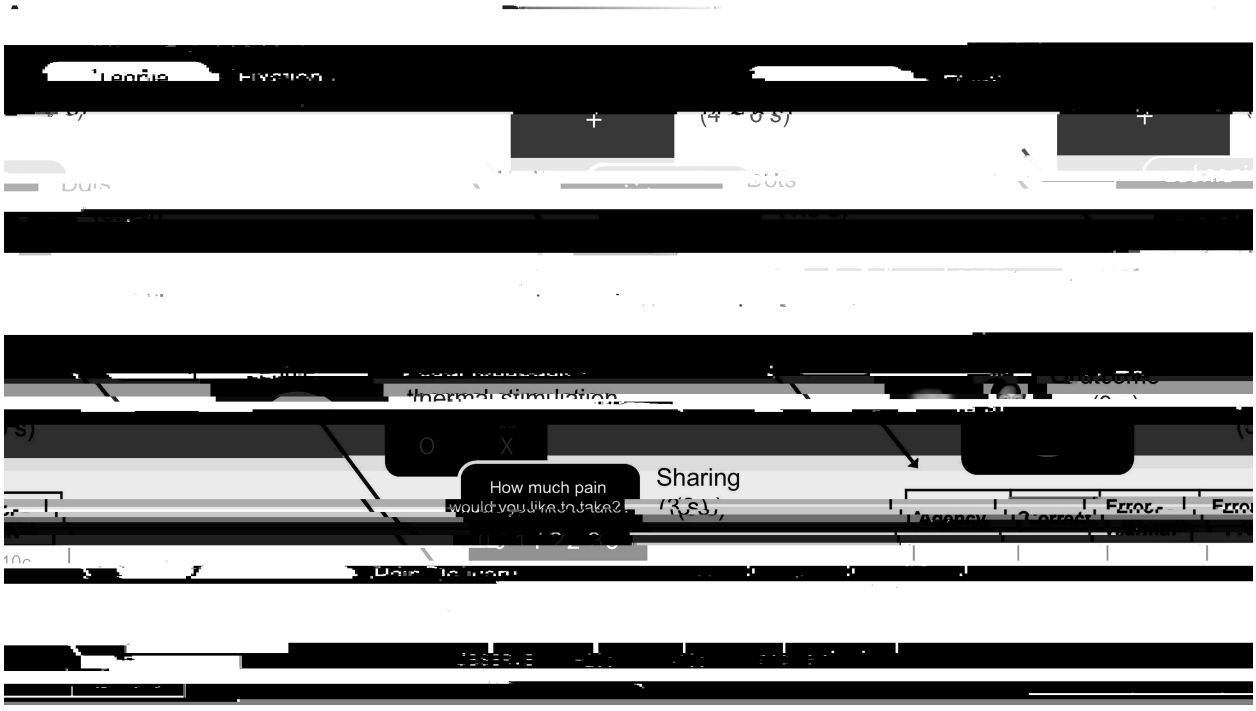

Figure 1. Procedure for Study 1 and Study 2. (*A*) In Study 1, the participant in the scanner was randomly paired with an anonymous partner on each trial. The task for the participant and the partner was to quickly estimate the number of dots presented briefly on the screen. The outcome of their performance was presented under the photo of the participant and under a blurred picture of face representing the partner. If at least one of them estimated incorrectly, the partner would receive a number of mildly painful electric shocks. The participant then indicated the level of pain he/she would be willing to take for the partner as a compensation. Finally, the pain stimulation of the participant's choice was delivered to him/her (see Yu et al. 2014 for details). (*B*

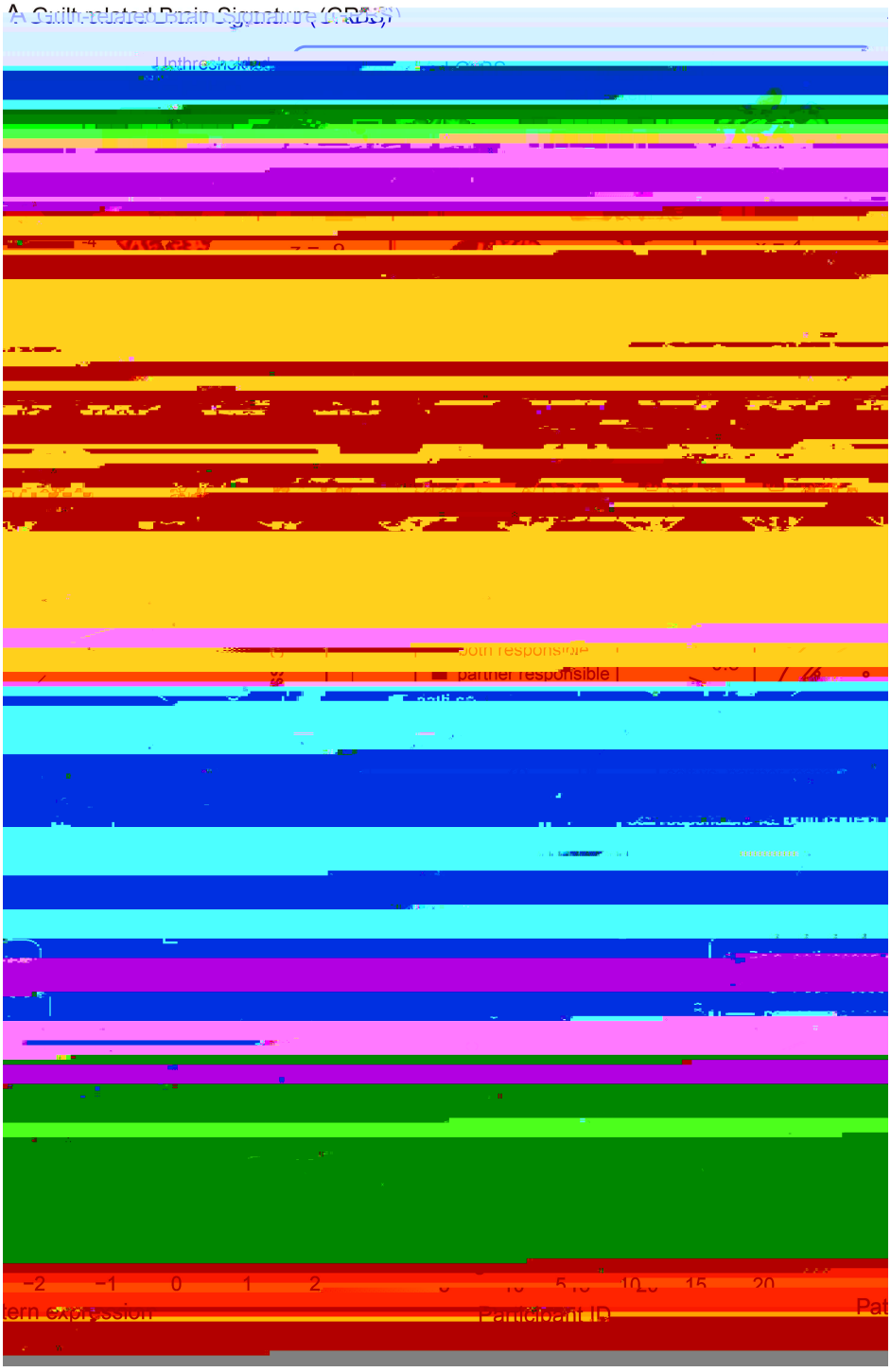

Figure 2. GRBS and its sensitivity. (*A*) Between-participant SVM weight map for guilt states (unthresholded). Bootstrap thresholded maps (5000 interactions, $z > 2$) is shown in the inset. Examples of unthresholded patterns within right insula (rAI) and anterior aMCC are also presented in the inset; small colored squares indicate voxel weights, black squares indicates empty voxels located outside of the GRBS pattern, and red-outlined squares indicate significance at $P < 0.005$ uncorrected (see also Table 1). (*B*) Cross-validated pattern expression computed as the dot product of the GRBS with the activation contrast maps for each participant. (*C*) ROC curves for the two-choice forced-alternative accuracies for the training data

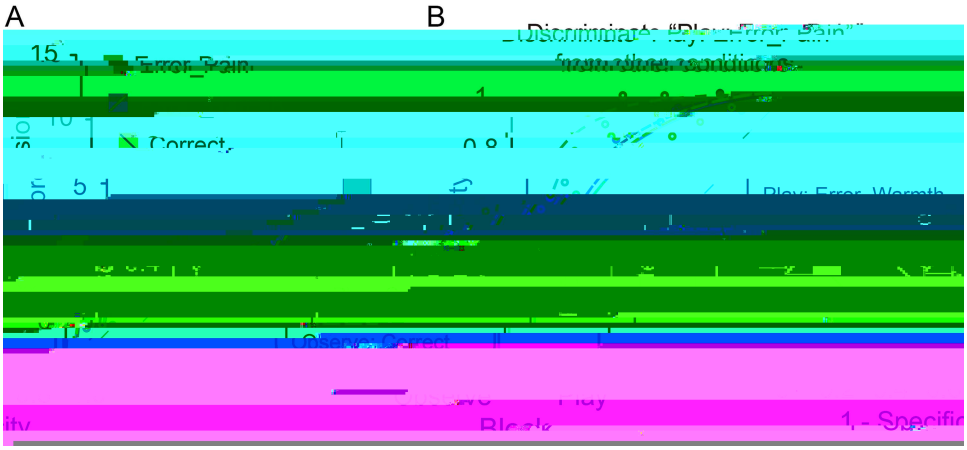

Figure 3. Generalizability of the GRBS. (*A*) In the Study 2 dataset, the "Play: Error_Pain" condition (i.e., the condition associated with highest guilt) shows the highest pattern expression. In this condition, the participant's action caused pain to the person outside the scanner (i.e., partner). In "Warmth" conditions, the participant's action may cause warm but not painful thermal stimulation to the partner. In "Correct" conditions, the participant did not make an error and no stimulation would be delivered to the partner. In "Observe" conditions, the participant observed the game and the pain stimulation was not contingent on their actions. Error bars indicate SEM. (*B*) ROC curves for the two-choice forced-alternative performance for the validation dataset (Study 2). Blue: Play Error Pain versus Play Error Warmth; Purple: Play Error Pain versus Observe Error Pain; Red: Play Error Pain versus Play Correct; Gold: Play Error Pain versus Observe Correct.

Testing the Specificity of the GRBS

To assess the specificity of the classifier, we examined its predictive power in two other independent data sets: one using thermal (heat) pain and observed (vicarious) pain (Krishnan et al. 2016), the other using recall task to elicit basic and social emotions (Wagner et al. 2011). Univariate analyses reported in these previous studies have implicated the brain regions showing highest predictive weights in the GRBS (e.g., aMCC, rAI) in the processing of physical and vicarious pain, and in the processing of recalled guilt episodes. However, it is an open question whether these brain states are distinguishable to GRBS. The multivariate approach allows us to test whether shared univariate activations reflect common neural representations (Woo et al. 2014). As can be seen from Figure 4 (see also Supplementary Table S3), GRBS performed at chance level in discriminating different intensity of thermal pain stimulation (High vs. Medium: accuracy = 57 ± 11%, $P = 0.57$; Medium vs. Low: accuracy = 46 ± 9%, $P = 0.85$) and different degree of vicarious pain (High vs. Medium: accuracy = 50 ± 9%, $P > 0.99$; Medium vs. Low: accuracy = 57 ± 9%, $P = 0.57$). The classifier did not significantly differentiate recalled guilt from either recalled sad memories (accuracy = 33 ± 12%, $P = 0.30$) or recalled shame memories (accuracy = 60 ± 13%, $P = 0.61$). These findings suggest that GRBS is better at detecting transgression in real-time interpersonal contexts than other unpleasant experiences, including guilt-related memories. That is, it does not appear to be selectively activated during retrieval of guilt-related memories, but it does respond selectively to feedback indicating that one has caused harm to a partner and predicts atonement behavior.

Finally, we investigated the relationship of the GRBS to other, potentially similar brain signatures of social-affective processes. Spatial similarity (Pearson correlation coefficients across all voxels) between the GRBS and eight other brain signatures related to social-affective processes are shown in Supplementary Table S4 and Figure S2. Most patterns showed around zero correlation (r's between −0.1 and 0.1), with the exception of the PINES—developed to track negative affect associated with unpleasant images (Chang, et al. 2015)—, which showed a weak positive correlation ($r = 0.12$) with GRBS, thus suggesting some shared variance between those two brain patterns. To examine this similarity more closely, we qualitatively examined whether it might be driven by shared positive or negative weights in ACC or insula, or other areas often activated by emotional events, such as the amygdala (ROIs defined based on anatomical labels and the WFU Pickatlas version 3.0.5b (Maldjian et al. 2003)). Figure 5*A* shows the joint distribution of normalized (z-scored) voxel weights of PINES on the x-axis and GRBS on the y-axis (cf. Koban et al. 2019). Differently colored octants indicate voxels of shared positive or shared negative (Octants 2 and 6, respectively), selectively positive weights for GRBS (Octant 1) and for PINES (Octant 3), selectively negative weights for GRBS (Octant 5) and for PINES (Octant 7), and voxels where the voxel weights of the two signatures went in opposite directions (Octants 4 and 8) (Fig. 5*B*). Overall correlations between the two patterns in the emotion mask (Fig. 5*A*) and in the three ROIs (Fig. 5*C–E*) were relatively weak. Across the whole emotion mask, stronger weights (sum of squared distances to the origin [SSDO]) were actually observed in the nonshared octants (1, 3, 5, 7). Further, the three ROIs showed distinct patterns of covariation between the two patterns. Many voxels in the bilateral amygdalae showed positive weights for PINES, but not for GRBS, as reflected by the high SSDO in Octant 3 (Fig. 5*C*). This is in line with the long-established role of the amygdala in emotional attention (see Vuilleumier 2005 for a review) and in assigning affective salience to sensory stimuli (LeDoux 2000). Bilateral insulae showed strongest weights in the Octants 1, 2, and 7, indicating many positive weights for guilt specifically (Octant1), as well as shared positive weights across the two signatures (Octant 2), but also some many voxels with negative weights in the PINES (Octants 6–8) (Fig. 5*D*). Finally, the ACC showed almost exclusively positive weights for GRBS,

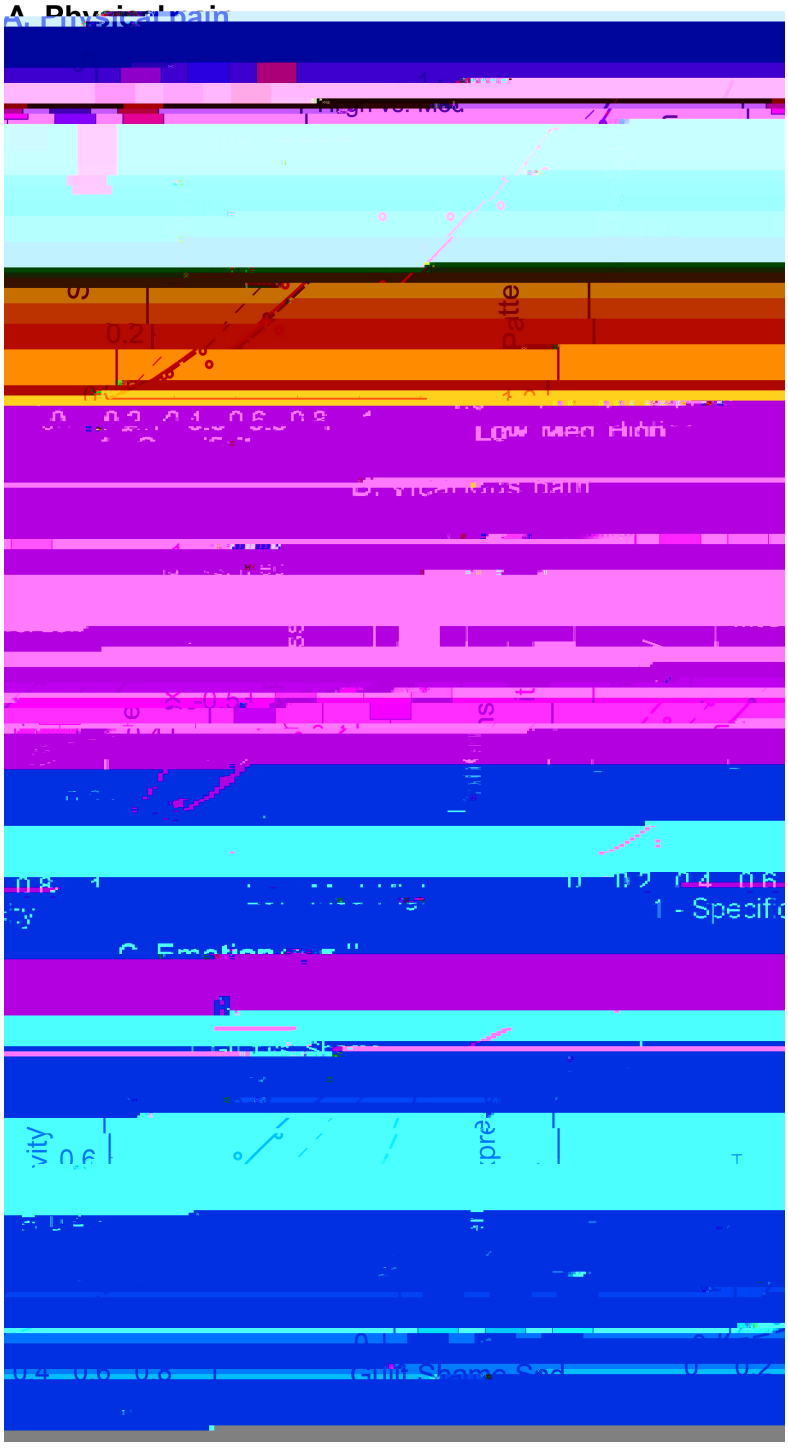

Figure 4. Specificity of the GRBS. (*A–C*

Hoffman ML. 2001. *Empathy and moral development: implications for caring and justice*. **Cambridge University Press.**

Hurtado de Mendoza A, Fernández-Dols JM, Parrott WG, Carrera P. 2010. Emotion terms, category structure, and the problem

Wager TD, Kang J, Johnson TD, Nichols TE, Satpute AB, Barrett LF. 2015. A Bayesian model of category-specific emotional brain responses. *PLoS Comput Biol*. **11:e1004066.**

Wagner U, N'Diaye K, Ethofer T, Vuilleumier P. 2011. Guilt-specific processing in the prefrontal cortex. *Cereb Cortex*. **21:2461–2470.**

Wong Y, Tsai J. 2007. Cultural models of shame and guilt. *The self-conscious emotions: Theory and research*. **209–223.**

Woo C-W, Koban L, Kross E, Lindquist MA, Banich MT, Ruzic L, Andrews-Hanna JR, Wager TD. 2014. Separate neural representations for physical pain and social rejection. *Nat Commun*. **5:5380.**

Woo C-W, Wager TD. 2015. Neuroimaging-based biomarker discovery and validation. *Pain*. **156:1379.**

Woo C-W, Chang LJ, Lindquist MA, Wager TD. 2017. Building better biomarkers: brain models in translational neuroimaging. *Nat Neurosci*. **20:365.**

Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD. 2011. Large-scale automated synthesis of human functional neuroimaging data. *Nat Methods*. **8:665.**

Yeung N, Nystrom LE, Aronson JA, Cohen JD. 2006. Between-task competition and cognitive control in task switching. *J Neurosci*. **26:1429–1438.**

Yu H, Duan Y, Zhou X. 2017. Guilt in the eyes: eye movement and physiological evidence for guilt-induced social avoidance. *J Exp Soc Psychol*. **71:128–137.**

Yu H, Hu J, Hu L, Zhou X. 2014. The voice of conscience: neural bases of interpersonal guilt and compensation. *Soc Cogn Affect Neurosci*. **9:1150–1158.**

Yu H, Siegel JZ, Crockett MJ. 2019. Modeling morality in 3-D: decision-making, judgment, and inference. *Topics Cognit Sci*. **11:409–432.**

Zahn-Waxler C, Kochanska G. 1990. The origins of guilt. In: Thompson RA, editor. *Nebraska symposium on motivation*. **Vol 36. Lincoln: University of Nebraska Press, pp. 183–258.**

Zhu R, Feng C, Zhang S, Mai X, Liu C. 2019. Differentiating guilt and shame in an interpersonal context with univariate activation and multivariate pattern analyses. *NeuroImage*. **186:476–486.**