eLIFE elifesciences.org

RESEARCH ARTICLE

Figures and figure supplements

Instructed knowledge shapes feedback-driven aversive learning in striatum and orbitofrontal cortex, but not the amygdala

Lauren Y Atlas *et al*

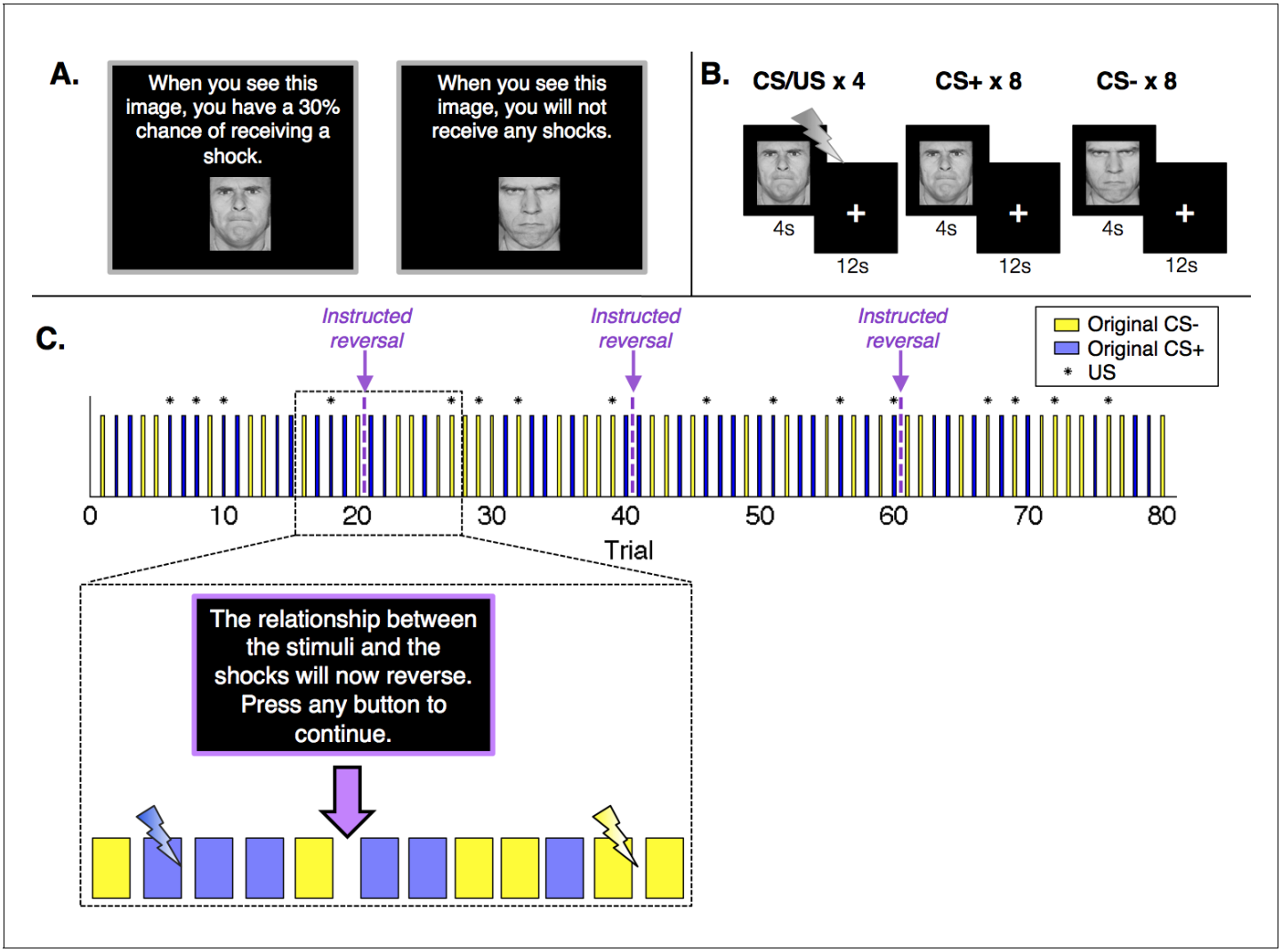

Figure 1. Experimental design. (**A**) Prior to the conditioning phase of the experiment, participants in the Instructed Group saw each image and were informed about initial probabilities. Participants in the Uninstructed Group also saw the images prior to the experiment, but were not told about contingencies. (**B**) Participants in both groups underwent a Pavlovian fear conditioning task with serial reversals. There were three reversals across the duration of the task, leading to four continuous blocks of twenty trials. In each block, one image (the conditioned stimulus, or CS+) was paired with a shock (the unconditioned stimulus, or US) 30% of the time, leading to 4 reinforced trials and 8 unreinforced trials, whereas a second image (the CS-) was never paired with a shock. Images were presented for 4 s, followed by a 12-second inter-stimulus interval. (**C**) Upon each reversal, the Instructed Group was informed that contingencies had reversed. Button presses were included to ensure participants were paying attention to the instructions but had no effect on the task itself or task timing. Instructions were always immediately followed by at least two unreinforced presentations of each CS before the new CS+ was paired with a shock. The figure presents one of two pseudorandom trial orders used during the experiment (see Materials and methods).

DOI: 10.7554/eLife.15192.003

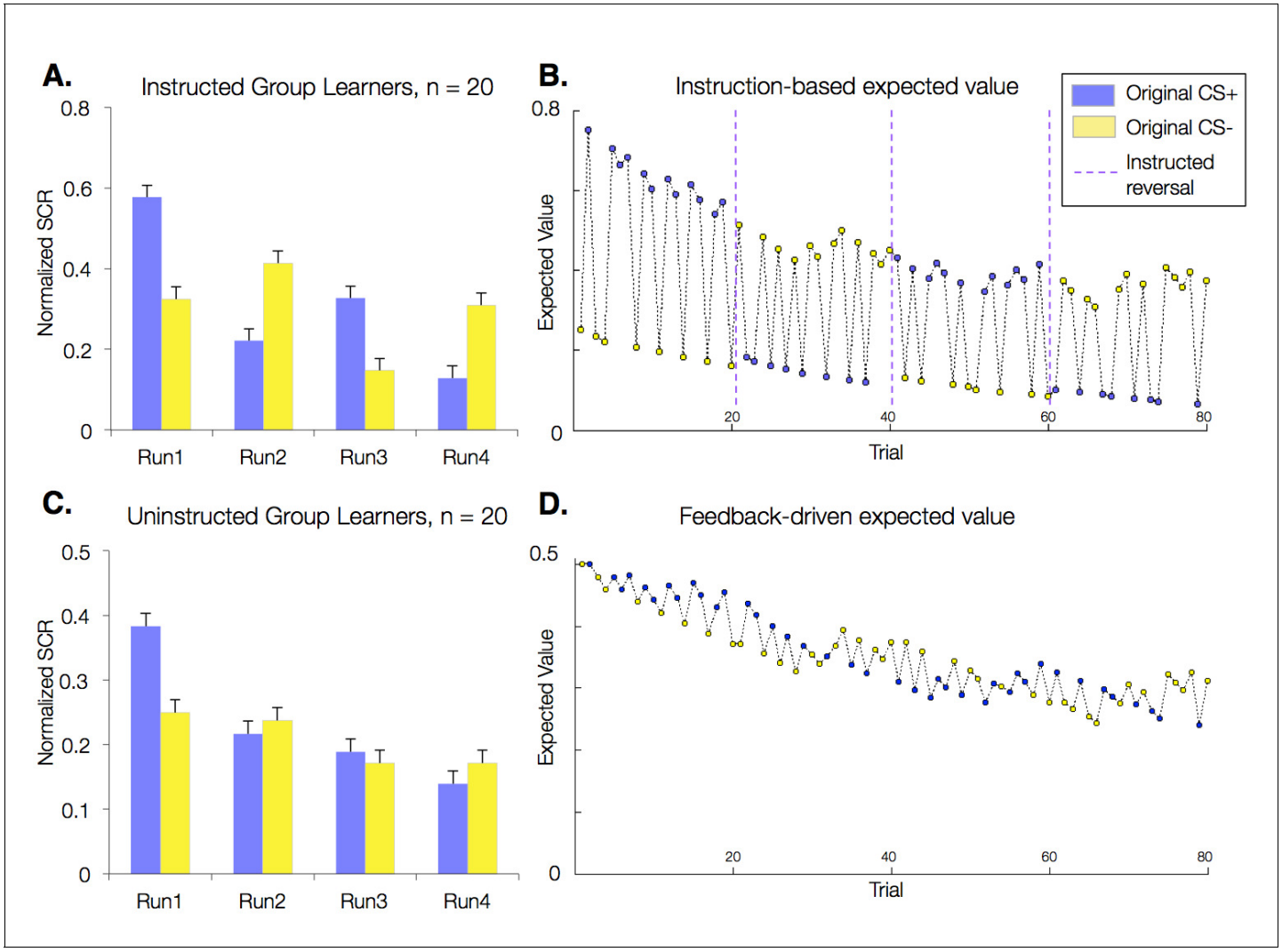

Figure 2. Effects of instructions on skin conductance responses (SCR) and aversive learning. Mean normalized skin conductance responses (SCRs) as a function of group and condition. Both groups showed significant reversals of SCR responses throughout the task ($p<0.001$; ***Table 1***), and effects were larger in the Instructed Group (see ***Table 1***). Error bars reflect within-subjects error. (**A**) Mean SCR in the Instructed Group as a function of original contingencies. Runs are defined relative to the delivery of instructions. (**B**) Dynamics of expected value based on fits of our modified Rescorla-Wagner model, fit to SCR in the Instructed Group. Fitted model parameters were consistent with SCR reversing almost entirely in response to instructions (ρ = 0.943). This timecourse was used in fMRI analyses to isolate regions involved in instruction-based learning. (**C**) Mean SCR in the Uninstructed Group as a function of original contingencies. A new run is defined when the previous CS- is paired with a shock. (**D**) Dynamics of expected value based on the model fit to SCR from the Uninstructed Group. This timecourse was used in fMRI analyses to isolate regions involved in feedback-driven learning in both groups.

DOI: 10.7554/eLife.15192.005

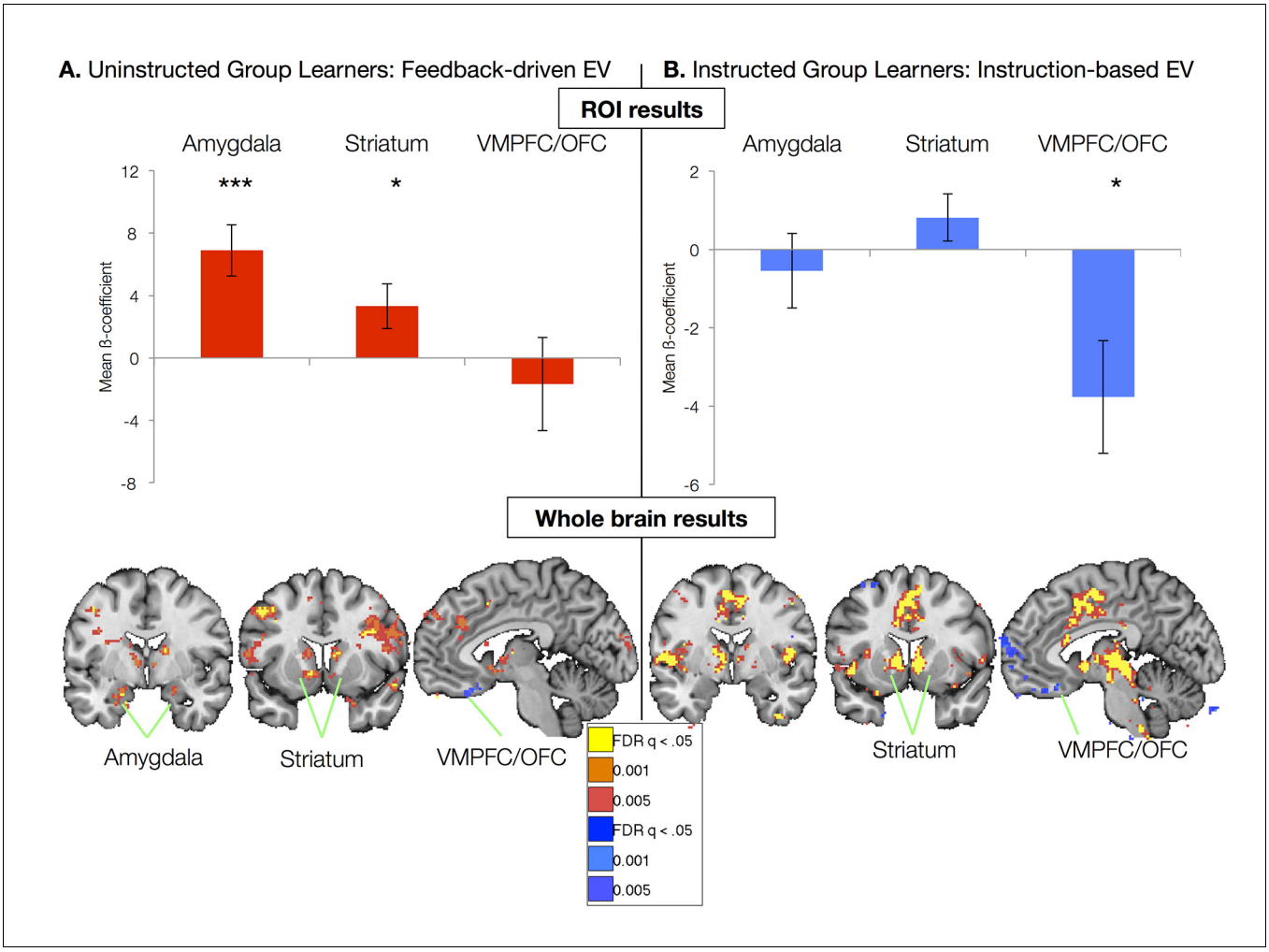

Figure 3. Neural correlates of expected value. (**A**) Neural correlates of feedback-driven expected value (EV) were isolated by examining correlations between the timecourse depicted in ***Figure 2D*** and brain activation in response to cue onset in Uninstructed Group learners (n = 20). *Top:* ROI-based analyses (see ***Figure 3—figure supplement 3***) revealed significant correlations with feedback-driven EV in the amygdala and striatum. Error bars reflect standard error of the mean; ***p<0.001; *p<0.05. *Bottom:* Voxel-wise FDR-corrected analyses confirmed ROI-based results and revealed additional correlations in the VMPFC/OFC, as well as other regions (see ***Figure 3—figure supplement 1***, ***Figure 3—figure supplement 1—source data 1***, ***2***). (**B**) Neural correlates of instruction-based EV were isolated by examining correlations between the timecourse depicted in ***Figure 2B*** and brain activation in response to cue onset in Instructed Group learners (n = 20). *Top:* ROI-based analyses revealed a significant negative correlation with instruction-based EV in the VMPFC/OFC. *Bottom:* Voxel-wise analyses confirmed these results and revealed strong positive correlations in the bilateral striatum, as well as the dACC, insula, and other regions (see ***Figure 3—figure supplement 2*** and ***Figure 3—figure supplement 2—source data 1*** and ***2***). We did not observe any correlations between amygdala activation and instruction-based EV.

DOI: 10.7554/eLife.15192.007

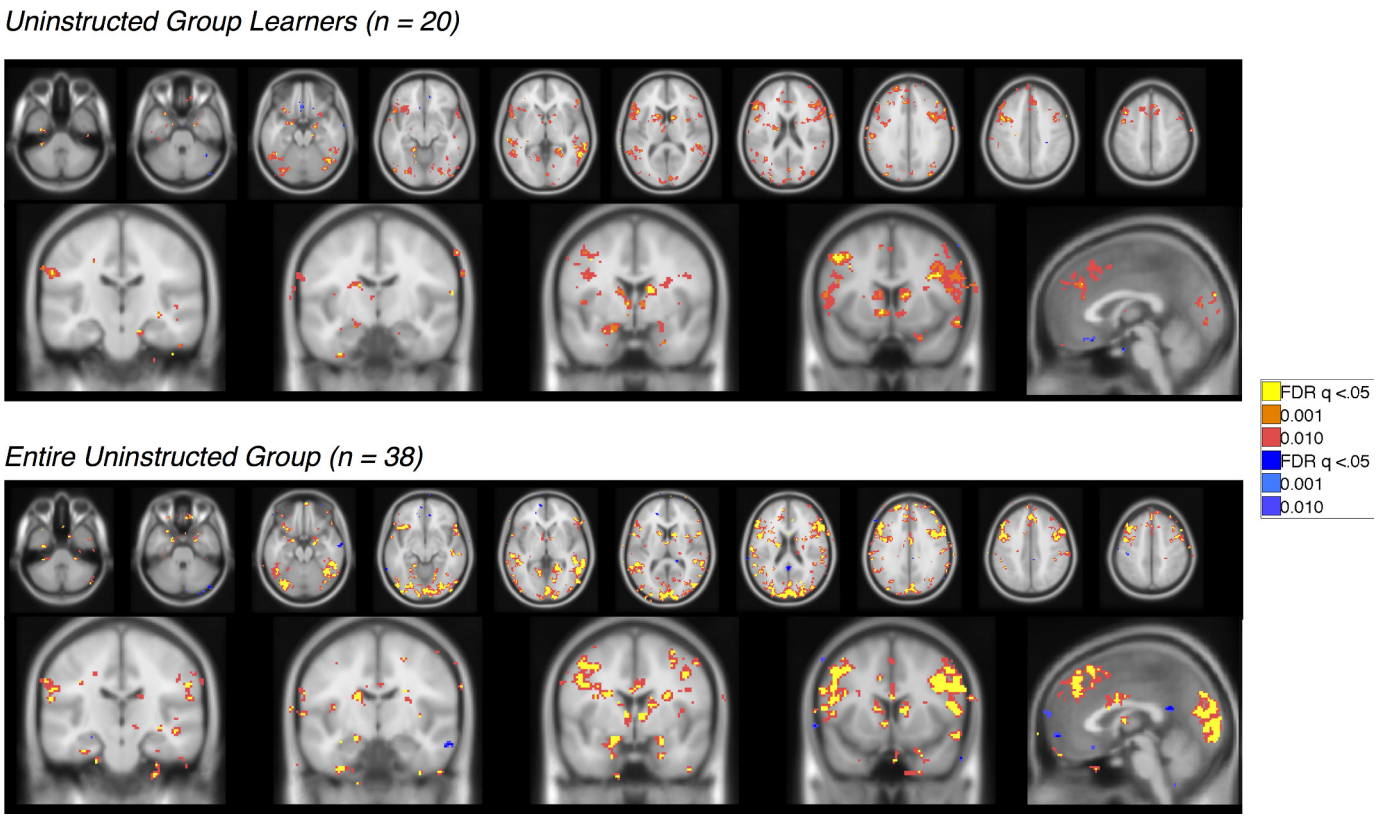

Figure 3—figure supplement 1. Feedback-driven EV in the Uninstructed Group. Voxelwise FDR-corrected results of neural correlates of feedback-driven EV in Uninstructed Group participants, based on the modified learning model fit to learners in the Uninstructed Group (across-subjects fits, see Materials and methods). Warm colors reflect positive correlations with EV, which was coded such that positive EV denotes expected shock. Cool colors reflect negative correlations. Top: Results in learners, or those individuals who showed differential SCR prior to the first reversal (n = 20). Bottom: Results across the entire Uninstructed Group (n = 38). For complete results in tabular format, please see '***Figure 3—figure supplement 1—source data 1***' and '***Figure 3—figure supplement 1—source data 2***'.

DOI: 10.7554/eLife.15192.008

The following source data is available for figure 3:

Figure supplement 1—Source data 1. Neural correlates of feedback-driven expected value (EV): Uninstructed Group Learners (n = 20).

DOI: 10.7554/eLife.15192.009

Figure supplement 1—Source data 2. Neural correlates of feedback-driven EV: Entire Uninstructed Group (n = 38).

DOI: 10.7554/eLife.15192.010

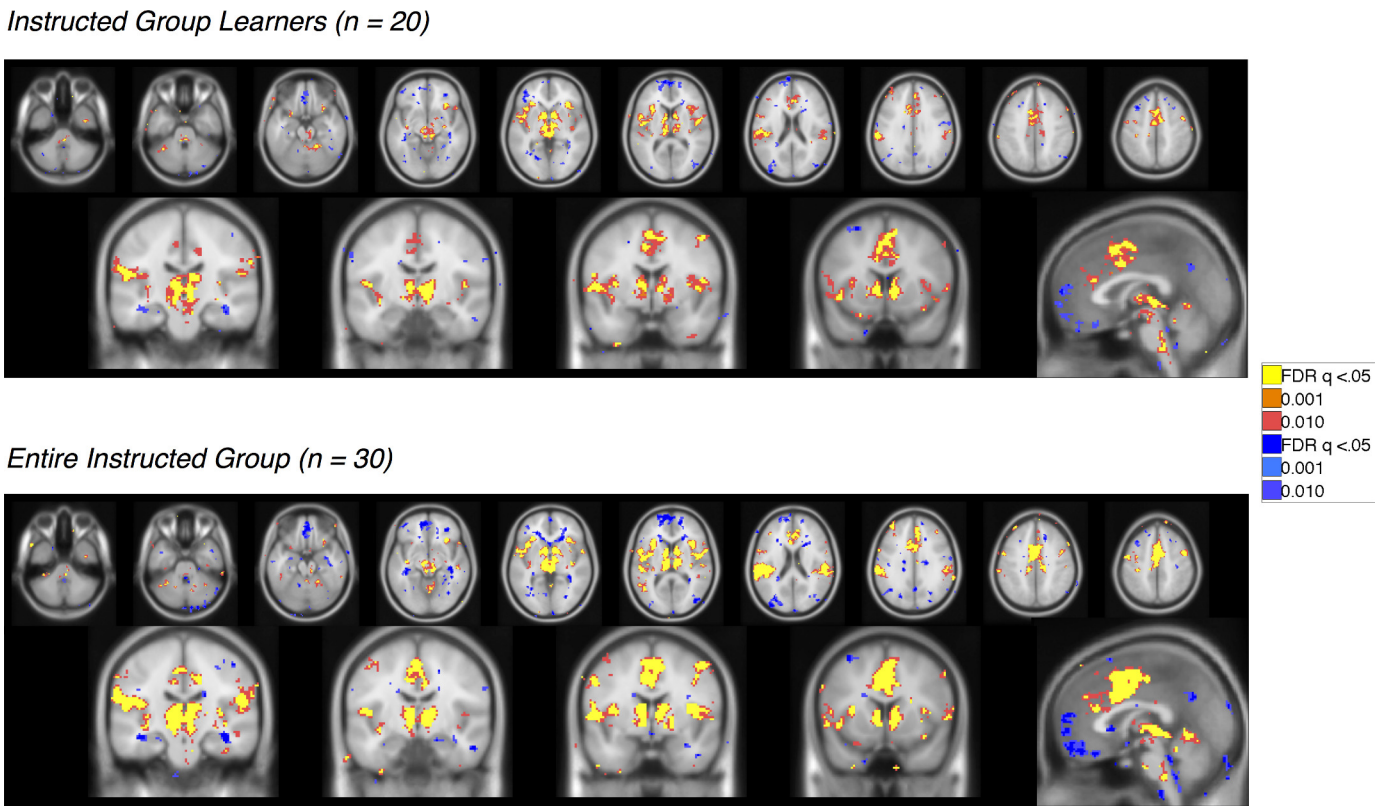

Figure 3—figure supplement 2. Instruction-based EV in the Instructed Group. Voxelwise FDR-corrected results of neural correlates of instruction-based EV in Instructed Group participants, based on the modified learning model fit to learners in the Instructed Group (across-subjects fits, see Materials and methods). Warm colors reflect positive correlations with EV, which was coded such that positive EV denotes expected shock. Cool colors reflect negative correlations. Top: Results in learners, or those individuals who showed differential SCR prior to the first reversal (n = 20). Bottom: Results across the entire Instructed Group (n = 30). For complete results in tabular format, please see '***Figure 3—figure supplement 2—source data 1***' and '***Figure 3—figure supplement 2—source data 2***'.

DOI: 10.7554/eLife.15192.011

The following source data is available for figure 3:

Figure supplement 2—Source data 1. Neural correlates of instruction-based EV: Instructed Group Learners (n = 20).

DOI: 10.7554/eLife.15192.012

Figure supplement 2—Source data 2. Neural correlates of instruction-based EV: Entire Instructed Group (n = 30).

DOI: 10.7554/eLife.15192.013

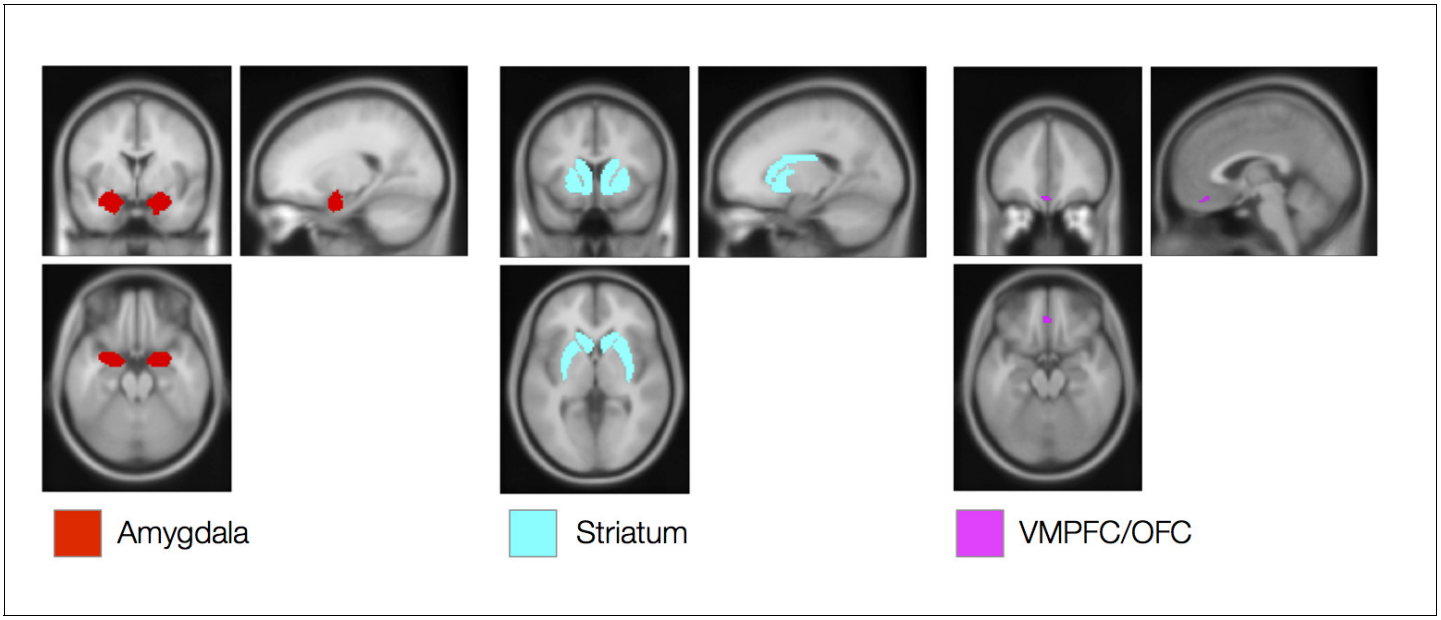

Figure 3—figure supplement 3. Regions of interest. Regions of interest for ROI-based FMRI analyses. See Materials and methods for details of ROI selection.

DOI: 10.7554/eLife.15192.014

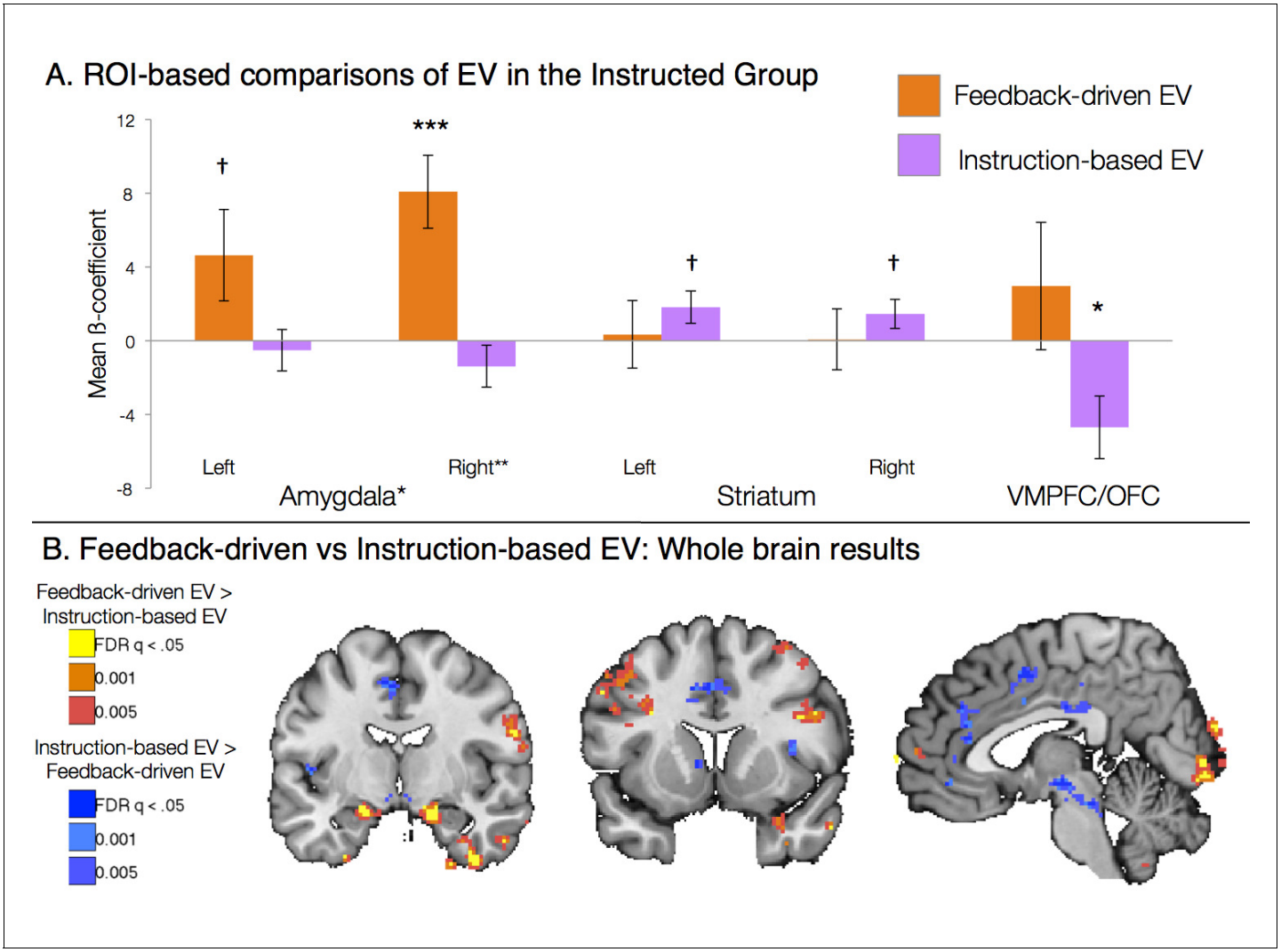

Figure 4. Dissociable effects of instructed and feedback-driven learning in the Instructed Group. (**A**) ROI-based effects of feedback-driven and instruction based EV signaling within Instructed Group learners from the model including both signals (see Materials and Methods). Direct model comparisons within Instructed Group learners revealed a significant effect of Model in the amygdala ($p<0.05$). VMPFC differences were marginal within learners ($p = 0.11$) and were significant when all Instructed Group participants were included in analyses ($p<0.05$). Error bars reflect standard error of the mean. ***$p<0.001$; *$p<0.05$; †$p<0.10$. (**B**) Voxelwise direct comparison between feedback-driven and instruction based EV signaling within the Instructed Group. Regions in warm colors, including bilateral amygdala (left), showed preferential correlations with feedback-driven EV. Regions in cool colors, including left caudate (middle), dorsal anterior cingulate and medial prefrontal cortex (right), showed higher correlations with instruction-based EV. Additional regions that showed significant differences as a function of model are presented in ***Figure 4—figure supplement 1***, ***Figure 4—figure supplement 1—source data 1*** and ***2***.

DOI: 10.7554/eLife.15192.015

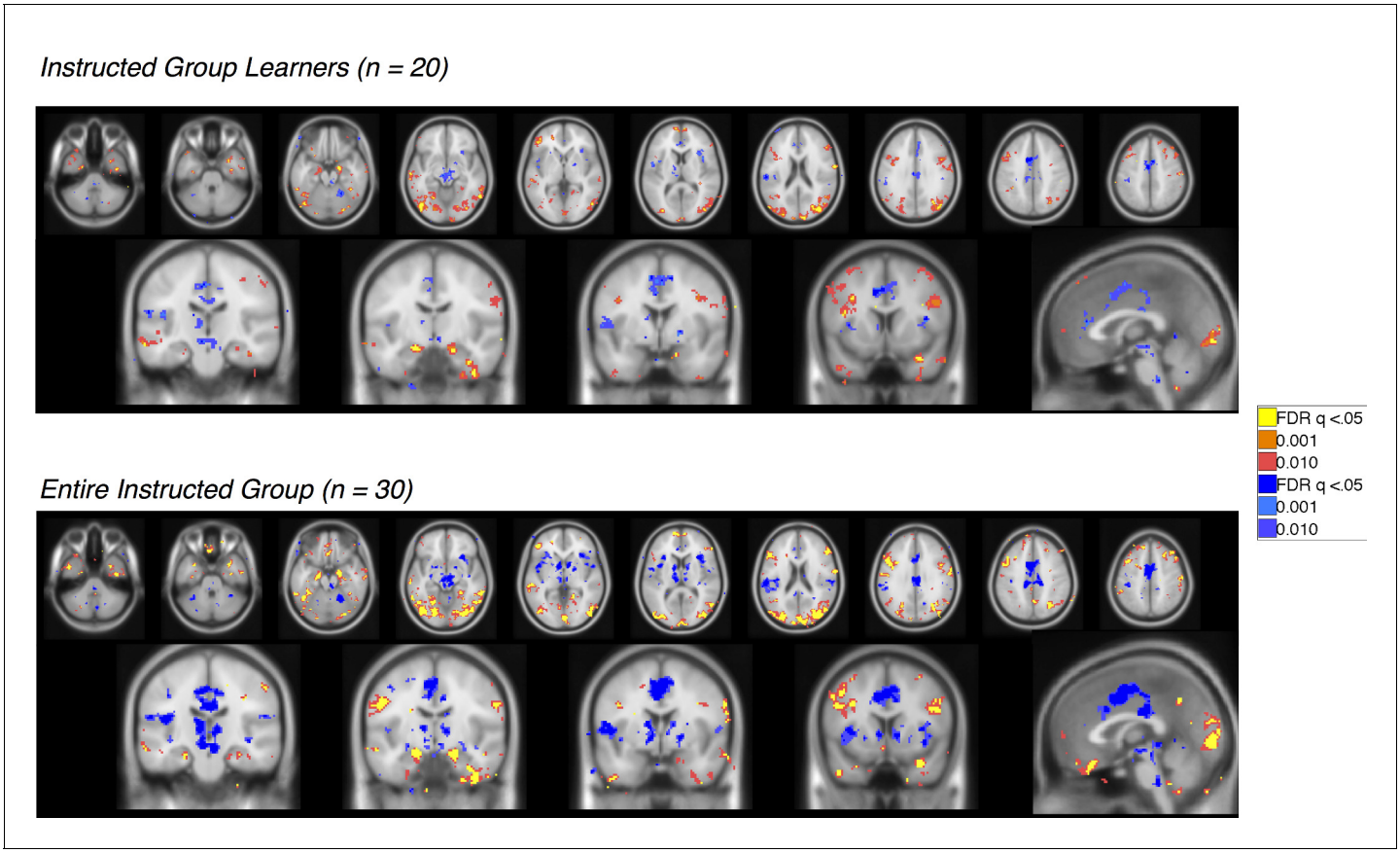

Figure 4—figure supplement 1. Feedback-driven vs instruction-based EV in the Instructed Group. Voxelwise FDR-corrected results of direct comparison between feedback-driven and instruction based EV signaling within the Instructed Group. Feedback-driven EV is based on the model fit to the Uninstructed Group learners, while instruction-based EV is based on fit to the Instructed Group learners. Regions in warm colors showed preferential correlations with feedback-driven EV, while regions in cool colors showed higher correlations with instruction-based EV. Top: Results in learners, or those individuals who showed differential SCR prior to the first reversal (n = 20). Bottom: Results across the entire Instructed Group (n = 30). For complete results in tabular format, please see '***Figure 4—figure supplement 1—source data 1***' and '***Figure 4—figure supplement 1—source data 2***.'

DOI: 10.7554/eLife.15192.016

The following source data is available for figure 4:

Figure supplement 1—Source data 1. Feedback-driven vs instruction-based EV: Instructed Group Learners (n = 20).

DOI: 10.7554/eLife.15192.017

Figure supplement 1—Source data 2. Neural correlates of instruction-based EV: Entire Instructed Group (n = 30).

DOI: 10.7554/eLife.15192.018

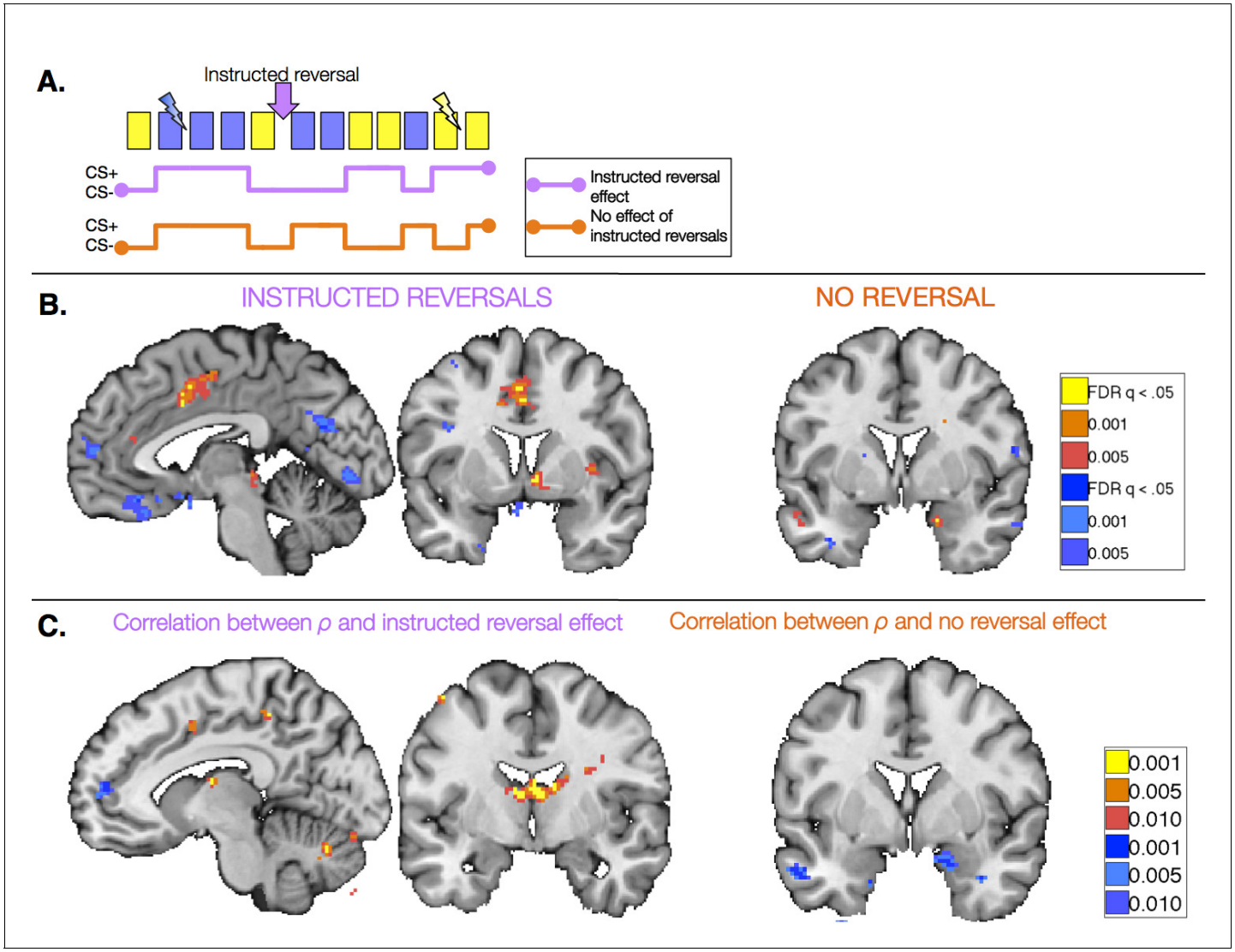

Figure 5. Task-based effects of instructed reversals and relationship with modeled behavior. (**A**) We examined responses in the Instructed Group surrounding the three instructed reversals to dissociate regions that are sensitive to instructions from those that learn from feedback and seem to be insensitive to instructions. Regions that are sensitive to instructions should show differential responses that reverse immediately upon instruction. The lavender timecourse depicts this pattern with greater activation on CS+ trials (blue) than CS- trials (yellow) prior to instruction, and the opposite pattern after instructions are delivered. Regions that update from aversive feedback and show no effect of instructed reversals would follow the orange timecourse, with greater activation to the previous CS+ than CS- both pre- and post-instruction. This feedback-driven pattern does not update until the new CS+ has been reinforced. (**B**) A number of regions showed differential responses that reversed upon instruction, including the right VS and VMPFC/OFC (left; see also ***Figure 5—figure supplement 1***, ***Figure 5—figure supplement 1—source data 1*** and ***2***). The VS showed greater activation to the current CS+ relative to the current CS-, whereas the VMPFC/OFC showed deactivation to the CS+. The right amygdala showed differential activation that did not reverse with instructions (right). Additional regions that did not reverse with instructions are presented in ***Figure 5—figure supplement 2***, ***Figure 5—figure supplement 2—source data 1*** and ***2***. (**C**) We conducted brain-behavior correlations to explore the relationship between neural activity in the period surrounding instructions and the magnitude of each individual's behavioral response to instructions. We tested for correlations between each individual's ρ parameter (based on within-subjects fits) and the magnitude of the reversal effect using an exploratory threshold of $p<0.001$, uncorrected. We observed significant correlations between ρ and the magnitude of instructed reversals in dACC (left) and the bilateral caudate tail and thalamus (right), as well as bilateral DLPFC (see ***Figure 5—figure supplement 3***), suggesting that those individuals who showed stronger reversals in SCR also showed stronger reversals in these regions. In addition, we found that the individuals who showed the least evidence for updating with instructions also showed the largest non-reversing differential responses in the right amygdala (right). Full results of brain-behavior correlations are reported in ***Figure 5—figure supplement 3*** and ***Figure 5—figure supplement 3—source data 1***.

DOI: 10.7554/eLife.15192.019

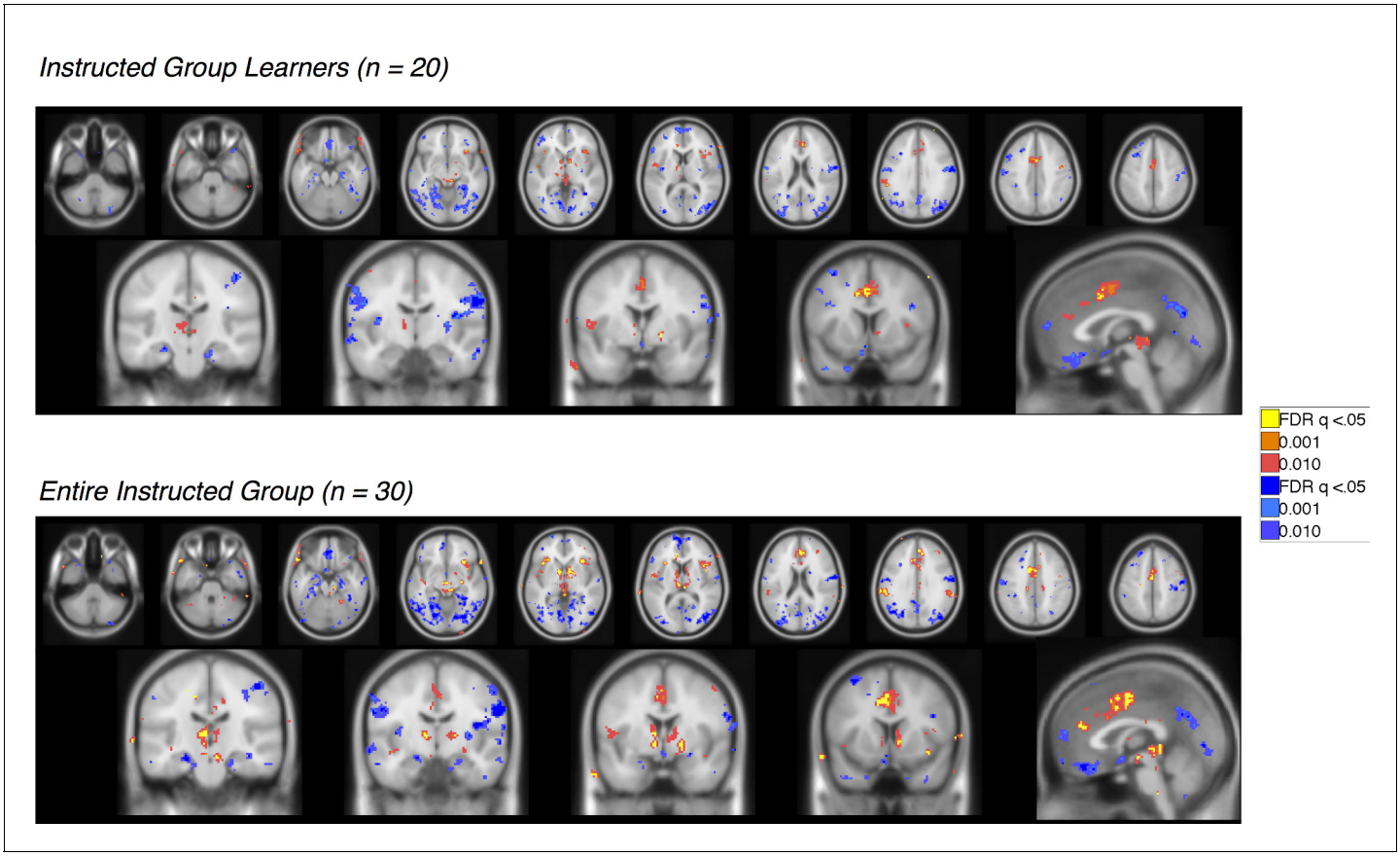

Figure 5—figure supplement 1. Immediate reversal with instructions (CS [previous CS+ > previous CS-] x Phase [Pre - Post] interaction) Voxelwise FDR-corrected results of regions that show immediate reversals with instructions in the Instructed Group, based on the window surrounding the delivery of instructions (see Materials and Methods). Regions in warm colors showed greater activation to the current CS+ relative to the current CS-, while regions in cool colors show relatively greater activation to the CS- (or deactivation to the CS+). Top: Results in learners, or those individuals who showed differential SCR prior to the first reversal (n = 20). Bottom: Results across the entire Instructed Group (n = 30). For complete results in tabular format, please see '***Figure 5—figure supplement 1—source data 1***' and '***Figure 5—figure supplement 1—source data 2***.'

DOI: 10.7554/eLife.15192.020

The following source data is available for figure 5:

Figure supplement 1—Source data 1. Immediate reversal with instructions (CS x Phase interaction): Instructed Group Learners (n = 20).

DOI: 10.7554/eLife.15192.021

Figure supplement 1—Source data 2. Immediate reversal with instructions (CS x Phase interaction): Entire Instructed Group (n = 30).

DOI: 10.7554/eLife.15192.022

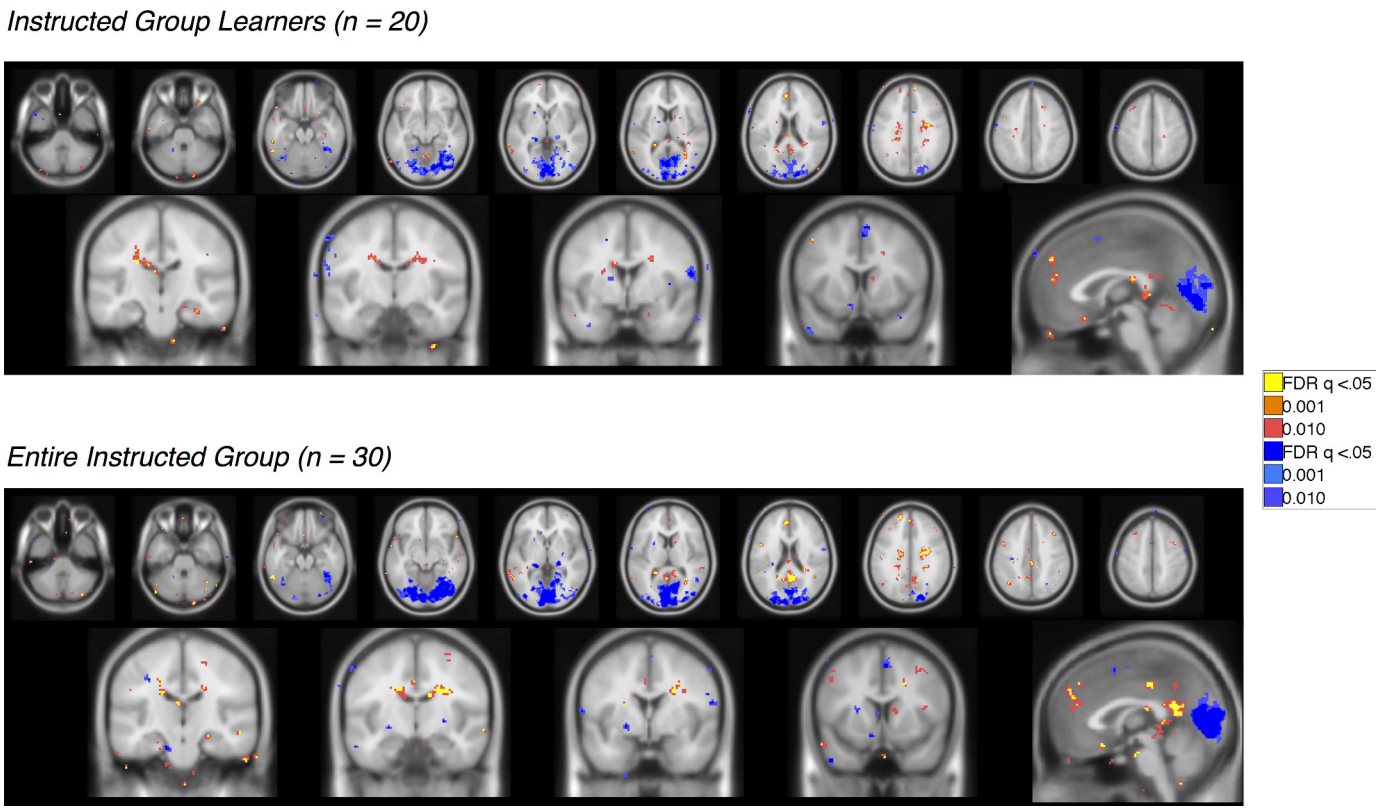

Figure 5—figure supplement 2. No reversal with instructions (main effect of CS without interaction; i.e. [previous CS+ > previous CS-] ⋃ [new CS- > new CS+]) Voxelwise FDR-corrected results of regions that show no evidence for reversal when instructions are delivered, based on continued differential responses pre- and post-instruction (see Materials and Methods). Regions in warm colors showed greater activation to the pre-instruction CS+ relative to the CS- both pre- and post-instruction, while regions in cool colors show relatively greater activation to the CS- (or deactivation to the CS+). Top: Results in learners, or those individuals who showed differential SCR prior to the first reversal (n = 20). Bottom: Results across the entire Instructed Group (n = 30). For complete results in tabular format, please see '***Figure 5—figure supplement 2—source data 1***' and '***Figure 5—figure supplement 2—source data 2***.'

DOI: 10.7554/eLife.15192.023

The following source data is available for figure 5:

Figure supplement 2—Source data 1. No reversal with instructions (main effect of CS without interaction): Instructed Group Learners (n = 20).

DOI: 10.7554/eLife.15192.024

Figure supplement 2—Source data 2. No reversal with instructions (main effect of CS without interaction): Entire Instructed Group (n = 30).

DOI: 10.7554/eLife.15192.025

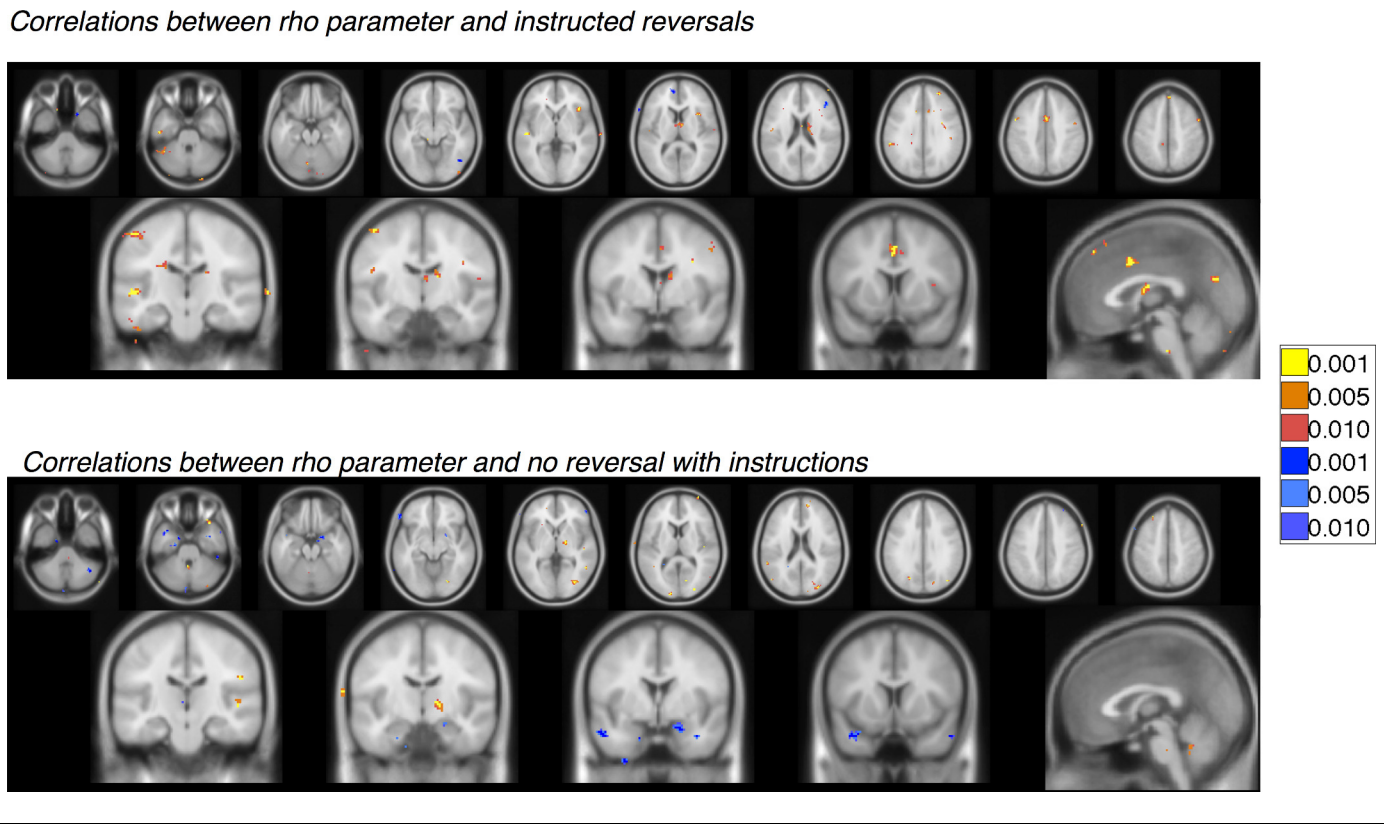

Figure 5—figure supplement 3. Correlations with instructed reversal (ρ) parameters. Voxelwise FDR-corrected results of brain-behavior correlations that tested for correlations between each individual's ρ parameter (based on within-subjects fits) and the magnitude of the reversal effect or absence of reversals using an exploratory threshold of $p<0.001$, uncorrected. Regions in warm colors show positive correlations between the magnitude of the instructed reversal parameter and the strength of reversal (Top) or sustained differential response (Bottom) in the region, while regions in cool colors show negative correlations. We focused on Instructed Group learners for these analyses. Top: Correlation between the instructed reversal parameter and immediate reversals with instructions (CS x Phase interactions). Bottom: Correlation between the instructed reversal parameter and the absence of instructed reversal (main effect of CS without reversal). For complete results in tabular format, please see '***Figure 5—figure supplement 3—source data 1***' and '***Figure 5—figure supplement 3—source data 2***.'

DOI: 10.7554/eLife.15192.026

The following source data is available for figure 5:

Figure supplement 3—Source data 1. Correlation between instructed reversal parameter (ρ) and instructed reversal effects: Instructed Group Learners (n = 20).

DOI: 10.7554/eLife.15192.027

Figure supplement 3—Source data 2. Correlation between instructed reversal parameter (ρ) and continued response to previous CS+ vs CS- (no reversal effect): Instructed Group Learners (n = 20).

DOI: 10.7554/eLife.15192.028

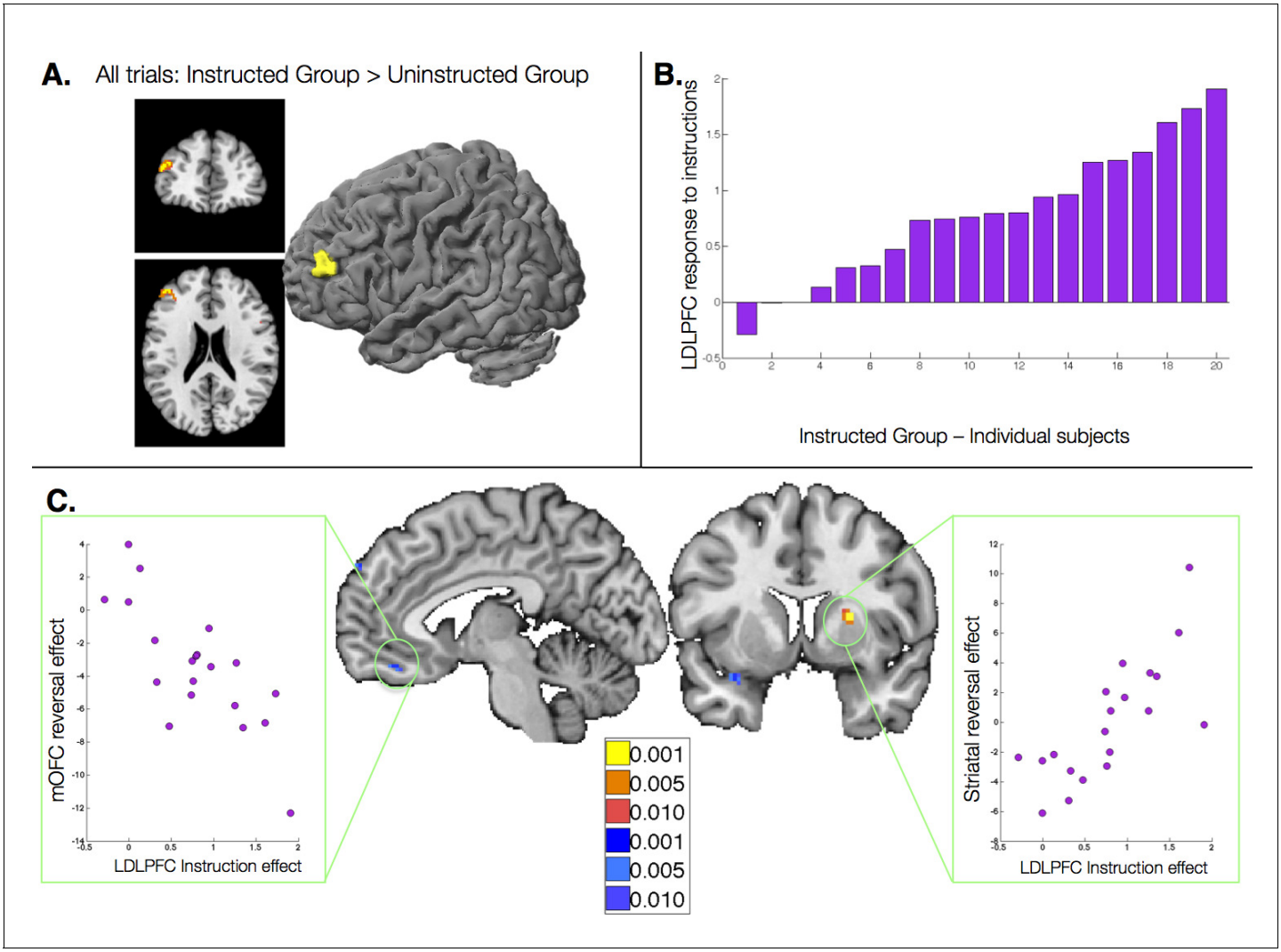

Figure 6. Relationship between dorsolateral prefrontal cortex response to instructions and instructed reversal effects. (**A**) The left dorsolateral prefrontal cortex (DLPFC) showed group differences across all trials, with greater activation in the Instructed Group than the Uninstructed Group. (**B**) We extracted the magnitude of the DLPFC response to instructions for each individual within the Instructed Group. **C**) The magnitude of the DLPFC response to instructions was correlated with the magnitude of instructed reversals in VMPFC/OFC (left) and dorsal putamen (right). High DLPFC responders showed larger reversals, with putamen activation to the new CS+ relative to the new CS-and VMPFC/OFC deactivation to the new CS+ relative to the new CS-. See also Figure Supplement 1 and ***Figure 6—figure supplement 1—source data 1***.

DOI: 10.7554/eLife.15192.029

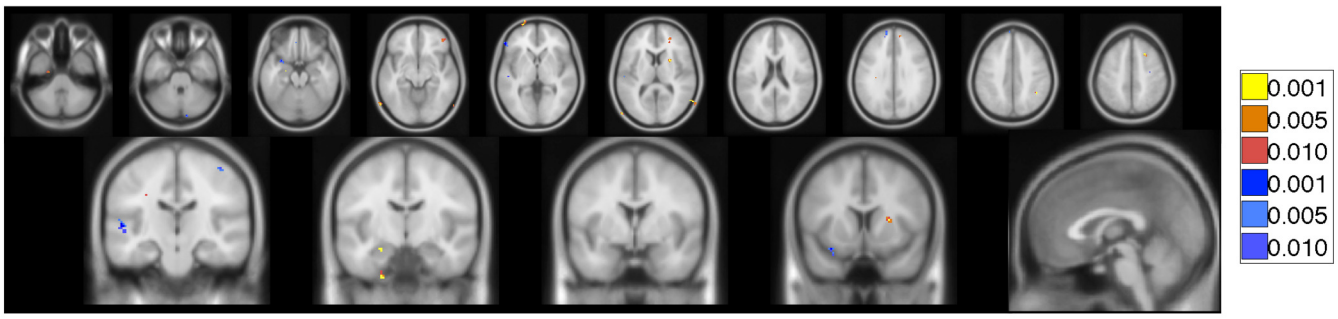

Figure 6—figure supplement 1. Correlations with left dorsolateral prefrontal cortex response to instructions. Voxelwise FDR-corrected results of brain-behavior correlations that tested for correlations between each Instructed individual's DLPFC response to instructions (see ***Figure 6***) and the magnitude of the reversal effect or absence of reversals using an exploratory threshold of $p<0.001$, uncorrected. Regions in warm colors show positive correlations between the magnitude of the DLPFC response and the strength of reversal, while regions in cool colors show negative correlations. We focused on Instructed Group learners for these analyses (n = 20). For complete results in tabular format, please see '***Figure 6—figure supplement 1—source data 1***'.

DOI: 10.7554/eLife.15192.030

The following source data is available for figure 6:

Figure supplement 1—Source data 1. Correlation between dorsolateral prefrontal response to instructions and instructed reversal effect: Instructed Group Learners (n = 20).

DOI: 10.7554/eLife.15192.031