NeuroImage 204 (2020) 116223

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Replication of fMRI group activations in the neuroimaging battery for the Mainz Resilience Project (MARP)

Miriam Kampa [a,b,*,1], Anita Schick [a,b,2], Alexandra Sebastian [b,c], Michèle Wessa [b,d], Oliver Tüscher [b,c], Raffael Kalisch [a,b,3], Kenneth Yuen [a,b,3]

[a] Neuroimaging Center (NIC), Focus Program Translational Neuroscience (FTN), Johannes Gutenberg University Medical Center, Mainz, Germany
[b] Deutsches Resilienz Zentrum (DRZ), Mainz, Germany
[c] Department of Psychiatry and Psychotherapy, Johannes Gutenberg University Medical Center, Mainz, Germany
[d] Department of Clinical Psychology and Neuropsychology, Institute of Psychology, Johannes Gutenberg University, Mainz, Germany

ARTICLE INFO

Keywords:
fMRI
Replication
Group activations
Jaccard index
Intra class correlation

ABSTRACT

Motivated by the recent replicability crisis we tested replicability of functional magnetic resonance imaging (fMRI) group activations in two independent samples. An identical behavioral and fMRI test battery for the longitudinal investigation of stress resilience mechanisms was developed for the Mainz Resilience Project (MARP) and conducted in a discovery (N = 54) and a replication sample (N = 103). The test battery consisted of a stress reactivity task, a reward sensitivity task, a fear conditioning and extinction paradigm, two volitional reappraisal tasks and an emotional interference inhibition task. Replicability of group activations was tested with the Jaccard index and the Intra Class Correlation (ICC). Overall, we observed good to excellent replicability of activations at the whole brain level. Only a minority of contrasts showed unsatisfactory replicability. Replicability at the level of individual regions of interest (ROIs) was generally lower. Tasks with stronger activation in the discovery sample showed better replicability.

1. Introduction

In their attempt to replicate the effects of 100 psychological experiments the Open Science Collaboration (2015) succeeded in only 40% of cases. Even lower replication rates of 11% have been reported in preclinical cancer research (Begley and Ellis, 2012). A subsequently conducted survey on the replication crisis in 1576 Nature readers revealed that a majority of scientists across disciplines has at least once experienced a failure in reproducing someone else's or even one's own results (Baker, 2016). Based on the current discussion, the conductance of replication studies is strongly recommended (Munafò et al., 2017). However, incentives for replication are low, since "systems in science favor novel findings over reliable ones" (Evans, 2017, p. 1; see also Button et al., 2013). This is especially true for neuroimaging research, where data acquisition is both costly and time-consuming (Turner et al., 2018).

Replication needs to be distinguished from reliability. Reliability indicates the precision of a measurement, commonly tested by the consistency across repeated measures (test-retest), whereas replication indicates the observation of a similar statistical pattern of results across different samples. Chances for replication decrease when error variance increases and reliability is reduced (De Schryver, Hughes, Rosseel and De Houwer, 2016; LeBel and Paunonen, 2011). Encouragingly, in recent years, there has been a growing body of functional magnetic resonance imaging (fMRI) studies estimating the test-retest reliability across different task domains (Caceres et al., 2009; Haller et al., 2018; Kristo et al., 2014; Lois et al., 2018; Moessnang et al., 2016; Nettekoven et al., 2018; Nord et al., 2017; Plichta et al., 2012; Quiton et al., 2014; Raemaekers et al., 2007). Test-retest reliability has been investigated for visual processing, motor tasks, emotional face processing, language, pain perception, working memory, receipt of reward and even complex mental processes like theory of mind. In contrast to former test-retest studies on group activations with repeated measurement of the same

* Corresponding author. Neuroimaging Center (NIC), Focus Program Translational Neuroscience (FTN), Johannes Gutenberg University Medical Center, Mainz, Germany.
E-mail address: miriam.kampa@psychol.uni-giessen.de (M. Kampa).
[1] Present address: Department of Clinical Psychology, University of Siegen, Germany. Bender Institute of Neuroimaging, Justus Liebig University Gießen, Germany.
[2] Present address: Public Mental Health, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany.
[3] These authors contributed equally.

https://doi.org/10.1016/j.neuroimage.2019.116223
Received 14 November 2018; Received in revised form 16 September 2019; Accepted 23 September 2019
Available online 23 September 2019
1053-8119/© 2019 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Abbreviations

Experimental conditions

Con Go	congruent go
CS	conditioned stimulus
CS+	CS that is reinforced during conditioning
CS+e	CS that is extinguished during extinction
CS+u	CS that is not extinguished during extinction
CS-	CS that is never reinforced
E-EX	early extinction (first two trials)
Incon go	incongruent go
L-EX	late extinction (last two trials)
Neg	negative
Neu	neutral
NT	no threat
NR	no reappraisal
Pos	positive
R	reappraisal
T	threat

Anatomical terms

ACC	anterior cingulate cortex
AI	anterior insula
dACC	dorsal anterior cingulate cortex
dlPFC	dorsolateral prefrontal cortex
IFG, p. tri.	inferior frontal gyrus - pars triangularis
NAcc	nucleus acccumbens
PFC	prefrontal cortex
pHPC	posterior hippocampus
OFC	orbitofrontal cortex
rdmPFC	rostral dorsomedial prefrontal cortex
vmPFC	ventromedial prefrontal cortex

sample (e.g. Caceres et al., 2009; Plichta et al., 2012), the replicability analysis performed here, compares group activations of two independent samples measured each at only one time point.

One of the rare studies quantifying the replicability of group activations in fMRI in a similar vein was conducted by Turner et al. (2018) who divided two big datasets into two random subsamples for a pseudo replication analysis. They found that replicability of group average activations increased with sample size. In conclusion, common fMRI studies with a typical sample size of N = 20 are underpowered, having an increased risk for missing true effects or producing false positives (Poldrack et al., 2017). Thus, using small sample sizes can actually contribute to generating irreproducible results (Turner et al., 2018).

Using two well-powered samples we here report about the replicability of fMRI group results from a combined behavioral and neuroimaging battery comprising several tasks that assess stress reactivity and regulation. A repeated employment of the battery is currently practiced in the prospective-longitudinal Mainz Resilience Project (MARP). In MARP, healthy young adults at risk for developing stress-related dysfunctions are being assessed regularly for exposure to negative life events and daily hassles as well as for potential changes in their mental health. These assessments are complemented by an extensive testing battery that serves to identify potential beneficial adaptations in biological, psychological and social functions that promote the maintenance of mental health despite adversity, i.e., resilience. A key element of the testing battery is a combined behavioral and neuroimaging battery that investigates the following functions: stress reactivity and recovery (in behavior and physiology only), reward sensitivity, safety learning and memory by means of Pavlovian fear conditioning and extinction, self-focused and situation-focused volitional reappraisal, response inhibition as well as emotional interference inhibition (all in fMRI), complemented by resting-state functional as well as structural and diffusion MRI scans (Kampa et al., 2018). In 2015, we initially tested the feasibility of the behavioral and imaging battery in a sample of N = 54 healthy young adults. In this sample (discovery sample), we observed that task-based activations conceptually replicate results from previous studies with similar tasks, as derived from meta-analyses, reviews or single studies, to satisfactory levels (Kampa et al., 2018). The MARP longitudinal study was started in 2016 and is still ongoing. For the replication analysis presented in this manuscript we used a subsample of 103 MARP subjects at their study entrance (replication sample). The goal of the analyses is to establish the robustness of the test battery as a preparatory step for the investigation of inter-individual differences and longitudinal changes using the test battery in a later stage.

2. Materials and methods

2.1. Subjects

All subjects were Caucasian with normal or corrected-to-normal vision. Included subjects were free of current psychiatric or neurological conditions and did not currently take any neuro-pharmacological or psycho-pharmacological medication. All subjects in the discovery sample and 98.1% in the replication sample were right-handed. All subjects gave written informed consent. Both studies (discovery and replication) were approved by the local ethics committee (Medical Board of Rhineland-Palatinate, Mainz, Germany) and conducted in accordance with the Declaration of Helsinki.

The discovery sample consisted of N = 54 subjects (30 women) and had a mean age of 25 years (age range: 18–31 years). The majority of the sample were university students (72.2%), one subject was still in secondary school (1.9%), 13.0% had a full-time job and 3.7% were unemployed. Overall, 53.7% of subjects (29 of 54) in the discovery sample had either a part- or full-time job aside from their studies. In the replication sample subjects were selected based on their former life history involving three or more negative life events. At study entrance, mental health was screened with the Mini International Neuropsychiatric Interview M.I.N.I. (Sheehan et al., 1998). The replication sample consisted of N = 103 subjects (64 women) with a mean age of 19 years (age range: 18–21 years). The majority of subjects in the replication sample were either university students (69.9%) or students in secondary school (8.7%); only two subjects were full-time employed (1.9%). The rate of subjects with a part- or full-time job was 36.8% (38 of 103) in the replication sample and thus significantly lower than in the discovery sample ($\chi 2 = 4.09$, $p = 0.04$). A comparison of the two samples concerning further characteristics including mental health, personality traits, social support and social status is given in Table 1. Bonferroni correction for multiple testing was applied to significant p-values by multiplication with the number of tests (34). The replication sample was significantly younger, had experienced more negative life events and scored lower on the coping strategy self-distraction.

2.2. Test battery

The combined behavioral and neuroimaging battery consisted of one behavioral and five fMRI tasks as well as of resting-state fMRI and structural and diffusion MRI scans. We will briefly introduce the behavioral and fMRI tasks, for which we tested replicability in the

Table 1
Descriptive statistics for discovery and replication sample.

Measure	Discovery			Replication			χ^2/t-score	p(p adjusted)
	N	*M*	*SD*	N	*M*	*SD*		
N (N women)	54(30)			103 (64)			0.64	0.42
Age	54	24.7	3.1	103	19.2	0.8	17.0	<0.001* (<0.001)
Negative life events[a]	54	6.8	2.7	101	9.7	4.5	−4.3	<0.001* (0.001)
Mental health (GHQ-28)[b]	54	48.5	8.41	100	50.0	10.5	−0.9	0.4
Anxiety sensitivity (ASI-3)[c]	53	19.1	10.5	101	15.4	10.2	2.1	0.035 (1.00)
Brief Resilience Scale (BRS)[d]	54	3.7	0.8	102	3.6	0.7	0.7	0.5
Optimism (LOT-R)[e]	54	15.4	4.5	102	16.0	4.2	−0.9	0.4
Self-efficacy (SWE)[f]	54	30.1	4.2	102	29.9	4.3	0.2	0.9
Coping style (BriefCOPE)[g]	54			102				
Self-distraction		5.6	1.3		4.8	1.4	3.6	<0.001* (0.01)
Denial		3.2	1.3		2.7	1.1	2.3	0.022
Emotional support		5.9	1.5		6.0	1.6	−0.4	0.7
Behavioral Disengagement		3.0	0.9		2.6	0.9	2.5	0.015 (0.51)
Positive reframing		5.2	1.6		5.1	1.6	0.3	0.7
Humor		4.2	1.8		4.2	1.7	−0.1	0.9
Active coping		5.7	1.3		5.3	1.4	1.6	0.1
Substance use		5.9	1.3		2.5	0.9	−0.5	0.6
Instrumental support		5.6	1.6		5.3	1.8	0.9	0.4
Living out emotions		4.3	1.3		4.3	1.5	0.0	1.0
Planning		5.9	1.3		5.7	1.4	0.9	0.3
Acceptance		6.0	1.5		5.5	1.7	1.9	0.1
Self-blame		4.4	1.6		4.1	1.4	1.4	0.2
Religion		3.1	1.7		2.9	1.5	0.7	0.5
Cognitive emotion regulation (CERQ)[h]	54			102				
Acceptance		7.7	1.9		7.8	2.0	−0.2	0.8
Rumination		6.2	2.0		6.0	2.0	0.4	0.7
Positive reappraisal		7.3	2.0		7.7	2.1	−1.2	0.2
Self-blame		4.8	2.1		5.2	2.0	−1.3	0.2
Positive refocusing		4.7	1.7		5.2	2.1	−1.6	0.1
Catastrophizing		4.4	2.3		4.2	1.9	0.5	0.6
Blaming others		3.6	1.6		3.6	1.4	0.1	0.9
Planning		7.1	1.8		7.2	2.1	−0.1	0.9
Putting into perspective		6.7	2.1		6.8	2.1	−0.4	0.7
Distancing[i]x		5.3	2.1		5.9	2.0	−1.7	0.1
Social support (OSS-3)[j]	54	10.6	1.8	76	10.8	1.9	−0.6	0.6
Subjective social status[k]								
Country	54	6.5	1.4	102	6.7	1.3	−1.2	0.2
Community	53	7.3	1.3	102	7.3	1.4	−0.1	0.9

*significant after Bonferroni-correction.
Bonferroni correction for multiple testing was applied to significant p-values. Adjusted p-values were computed by multiplying the original p-value with the number of tests (34). M, mean; SD, standard deviation.
[a] ad. from Caspi et al. (1996), translated to German; Likert scoring from 0 to 3.
[b] General health questionnaire GHQ-28; German version: Klaiberg et al. (2004); original: Goldberg and Hillier (1979).
[c] Anxiety sensitivity index ASI-3; Kemper et al. (2009); Taylor et al. (2007).
[d] Brief resilience scale BRS; Chmitorz et al. (2018); Smith et al. (2008).
[e] Life orientation test – revised LOT-r; Glaesmer et al. (2008); Scheier and Carver (1985).
[f] Selbstwirksamkeitserwartung SWE; Schwarzer and Jerusalem (1999).
[g] Brief coping inventory Brief COPE; Knoll et al. (2005); Carver (1997).
[h] Cognitive emotion regulation questionnaire CERQ; Loch et al. (2011); Garnefski et al. (2001).
[i] Two items reflecting distancing/detachment were added to the original questionnaire (English translation: "I try to observe the situation from a detached perspective, like from outside.", "I try to distance myself from the situation and my feelings.").
[j] Oslo social support scale OSS-3; ad. from Dalgard et al. (1995); translated to German.
[k] Euteneuer et al. (2014); Adler et al. (2000).

current report. For a detailed description of the procedure and methods, see Kampa et al. (2018) and the figures on the task designs in the supplementary materials S1–S6. The behavioral task (Task 1) was presented during visit 1, where subjects also filled in questionnaires and were instructed about the fMRI tasks. The fMRI tasks (Tasks 2–6) were presented during visits 2 and 3. Visits 1 and 2 took place maximally four weeks apart. Visit 3 always took place one day after visit 2. Table 2 gives an overview of the tasks and the tested outcomes. See suppl. materials S7 for a description of the acquisition and analysis of physiological and behavioral data.

Table 2
Overview of tasks and tested outcomes.

Task	Description	Tested outcomes
1	Stress reactivity and recovery	Stress ratings, heart rate, skin conductance, salivary cortisol and alpha amylase
2	Reward sensitivity	Reaction times, fMRI
3	Safety learning and memory	Fear ratings, skin conductance, fMRI
4	Self-focused volitional reappraisal	Fear ratings, skin conductance, fMRI
5	Situation-focused volitional reappraisal	Emotional state ratings, fMRI
6	Emotional interference and motor response inhibition	Reaction times, stop signal reaction times, fMRI

2.2.1. Task 1: Stress reactivity and recovery

Task 1 was conducted in the behavioral laboratory and tested stress reactivity and recovery by presenting multiple stressors on a screen and

via head phones (adapted version of the Mannheim Multicomponent Stress Test, MMST; Reinhardt et al., 2012). Since we did not acquire MRI during this task, we can not report on the replicability of group activations. For the interested reader details on the task are given in the suppl. materials (Fig. S1).

2.2.2. Task 2: Reward sensitivity

Task 2 assessed reward sensitivity using an adapted version of the monetary incentive delay task (MID; Knutson et al., 2001; Wu et al., 2014). See suppl. Fig. S2 for the design of Task 2. Each trial started with a 2-s cue indicating the incentive condition (+3 €, +0.5 €, ±0 €, −0.5 €, −3 €). The cue was followed by an anticipation phase of 2–2.5 s during which subjects had to press a button as soon as the target stimulus appeared on the screen, in order to gain or to avoid losing the indicated amount. Each trial ended with a 2-s numeric feedback on subjects' trial outcome and overall outcome. To assure that the experience of reward did not differ between subjects depending on task performance, an adaptive algorithm was applied that changed target duration for a given subject within each condition based on past performance. If the subjects' hit rate was below 66%, target duration was increased by 25 ms; else, it was reduced by 25 ms. Reaction times and hit rates were collected as behavioral outcomes. Results for reaction times are given in the suppl. materials (Table S1). For the fMRI replicability analyses, the main contrasts of interest were anticipation of gain (Anticipation: Gain > Zero; region of interest (ROI): right nucleus accumbens, R NAcc), anticipation of loss (Anticipation: Loss > Zero, ROI: R NAcc), gain outcome (Outcome: Gain > No Gain, ROI: bilateral ventromedial prefrontal cortex, B vmPFC) and no loss outcome (Outcome: No Loss > Loss, ROI: left nucleus accumbens, L NAcc). The ROIs selected for replicability testing in this and the other tasks (Fig. 1) were those with the strongest effect size (highest z-score) in a given contrast in the discovery sample (Kampa et al., 2018).

2.2.3. Task 3: Safety learning and memory

Task 3 investigated fear and safety learning by means of differential Pavlovian fear conditioning, extinction and a memory retrieval test that involved spontaneous recovery and renewal (Kalisch, 2006; Milad et al., 2007). See suppl. Fig. S3 for the design of Task 3. During Pavlovian fear conditioning in context A, two conditioned stimuli (CSs) were coupled with the unconditioned stimulus (US, painful electric stimulation) in 100% of trials, to become CS+s. A third CS (CS-) was never followed by a US and therefore safe. During subsequent extinction in context B, one of the CS+s (CS+e) and the CS- were presented in the absence of any US. The retrieval tests were then conducted on the following day. To test spontaneous recovery, all three CSs (CS+e, the unextinguished CS+u, and CS-) were presented in context B in the absence of a US; to test renewal, the same procedure was repeated in the original conditioning context A. Trial-by-trial fear ratings and skin conductance responses (SCR) served as behavioral outcome measures. For the fMRI replicability analyses, the main contrasts of interest for conditioning were the CS+s vs. CS- comparisons (CS+s > CS-, ROI: bilateral dorsal anterior cingulate cortex, B dACC; CS- > CS+s, ROI: B vmPFC), whereas in extinction we compared early vs. late responses to both the CS+e and the CS- combined (E-EX > L-EX, ROI: right anterior insula, R AI; L-EX > E-EX, ROI: right posterior hippocampus, R pHPC). In spontaneous recovery, we compared CS+u and CS- (CS+u > CS-, ROI: R AI; CS- > CS+, ROI: B vmPFC), assessing retrieval of unextinguished conditioned responses (to CS+u) relative to the safety signal established during conditioning (CS-), as well as CS+u > CS+e (ROI: R AI), assessing unextinguished conditioned responses relative to the new safety signal established in extinction (CS+e). In renewal, where return of fear was generalized (i.e., both to CS+u and CS+e), we compared CS+s and CS- (CS+s > CS-, ROI: L AI; CS- > CS+s, ROI: right orbitofrontal cortex, R OFC).

2.2.4. Task 4: Self-focused volitional reappraisal

Task 4 employed an adapted version of the instructed fear task used by Paret et al. (2011), in which fear was induced by telling subjects that one of two symbols indicated the occurrence of a painful electric stimulus with 25% probability (Threat condition, T), whereas the other symbol indicated safety (No Threat, NT). See suppl. Fig. S4 for the design of Task 4. When instructed, subjects had to volitionally reappraise (reduce) the self-relevance of the symbols using a distancing, or detachment strategy (Reappraisal condition, R; compared to a No Reappraisal condition, NR). The task involved a fully balanced, two-by-two factorial design with the experimental factors threat (T, NT) and reappraisal (R, NR). Fear ratings and skin conductance level (SCL) served as behavioral outcome measures. For the fMRI replicability analyses, the main contrasts of interest were the main effects of threat (T > NT, ROI: bilateral rostral dorsomedial prefrontal cortex, B rdmPFC; NT > T, ROI: B vmPFC) and reappraisal (R > NR, ROI: left dorsolateral prefrontal cortex, L dlPFC). Note that, for reasons of comparability with the analysis of Task 5, we selected a L dlPFC ROI for the reappraisal contrast, although it was only the second most activated region in this task (strongest activation in left supplementary motor area).

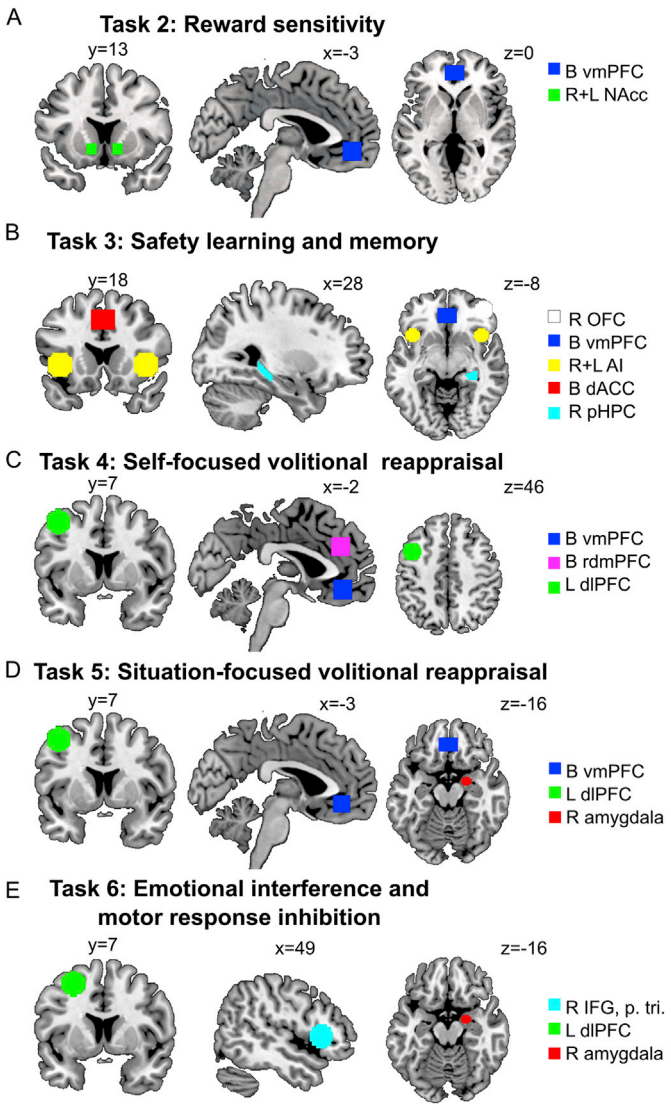

Fig. 1. Overview of regions of interest (ROIs) used in replicability analysis. B, bilateral; L, left; R, right. AI, anterior insula; dACC, dorsal anterior cingulate cortex; dlPFC, dorsolateral prefrontal cortex; IFG, p. tri., inferior frontal gyrus – pars triangularis; NAcc, nucleus accumbens; OFC, orbitofrontal cortex; pHPC, posterior hippocampus; rdmPFC, rostral dorsomedial prefrontal cortex; vmPFC, ventromedial prefrontal cortex.

2.2.5. Task 5: Situation-focused volitional reappraisal

In Task 5 subjects had to positively reinterpret either negative (Neg), positive (Pos) or neutral (Neu) pictures. The paradigm is an adaptation of Kanske et al. (2011). See suppl. Fig. S5 for the design of Task 5. We used a fully balanced, three-by-two factorial design, combining the three types of picture valences with either situation-focused reappraisal (R) or viewing the pictures as a control condition (No Reappraisal, NR). Emotional state ratings served as outcome measures. For the fMRI replicability analyses, the main contrasts of interest were the simple main effects of picture viewing (Pos/NR > Neu/NR, ROI: B vmPFC; Neg/NR > Neu/NR, ROI: R amygdala) and the main effect of reappraisal (R > NR, ROI: L dlPFC). Note that, for reasons of comparability with the analysis of Task 6, we selected a R amygdala ROI for the negative picture viewing contrast, although it was only the second most activated region in this task (strongest activation in left inferior lateral occipital cortex).

2.2.6. Task 6: Emotional interference and motor response inhibition

Task 6 tested the inhibition of the interference induced by emotional picture stimuli with performance on a motor response inhibition task (HRI, Hybrid Response Inhibition task; Sebastian et al., 2013). Two HRI-subtasks were included in Task 6: the Simon task, capturing spatial interference inhibition, and the stop signal task, assessing action cancellation, both in comparison to a simple go task as control condition. Stimuli were a fixation cross and an arrow on the left or right side of the cross, both encircled by a white ellipse. Subjects had to indicate the pointing direction of the arrow. In congruent go trials (Con Go), the position of the arrow relative to the cross corresponded to its pointing direction. In incongruent go trials (Incon Go), arrow position and pointing direction were opposite, inducing spatial interference. In stop trials, which were all congruent go trials, the white ellipse turned blue after a stop signal delay (SSD), requiring subjects to cancel their prepared or ongoing action. The SSD was adapted using a staircase tracking procedure (separately for Neg and Neu trials) based on past performance to achieve a correct stopping rate of 50%. The range of possible SSDs was between 30 and 540 ms. To additionally induce emotional interference with these two motor tasks, negative (Neg) and neutral (Neu) pictures were used as primes, shown in the 500 ms before a task trial. See suppl. Fig. S6 for the design of Task 6. Reaction times and stop signal reaction times served as the behavioral outcomes of interest. For the fMRI replicability analyses, the main contrasts of interest were response interference inhibition (Incon Go > Con Go, ROI: L dlPFC), stopping (Stop > Con Go, ROI: right inferior frontal gyrus, pars triangularis, R IFG, p. tri.), negative primes (Neg > Neu, ROI: R amygdala), and an interaction between emotion and motor response inhibition ((Con Go)>(Incon Go, Stop)/2)$_{Neg}$ > ((Con Go)>(Incon Go, Stop)/2)$_{Neu}$, ROI: R amygdala). Note that, for reasons of comparability with the analysis of Task 5, we selected a R amygdala ROI for the negative prime contrast, although it was only the second most activated region in this task (strongest activation in L amygdala).

2.3. MRI data acquisition

Images were acquired on a Siemens 3 T-Magnetom Tim Trio system (Siemens, Germany) running on software version Vb17, using a 32-channel head coil. Foam pads restricted head movement. Visual stimuli were presented on a screen at the head end of the scanner bore and projected to the subject's visual field via a mirror that was fixed on the head coil. A multiband echo planar imaging (EPI) sequence (TR = 1000 ms, TE = 29 ms, flip angle = 56°, FOV = 210 mm, voxel size = 2.5 × 2.5 × 2.5 mm^3, 60 slices, MB acceleration factor = 4, Bandwidth = 2588 Hz/px, no further GRAPPA acceleration) from the Center for Magnetic Resonance Research, University of Minnesota (CMRR) adopted from the Human Connectome Project was used for blood oxygen-level dependent (BOLD) (Feinberg et al., 2010) fMRI. fMRI was complemented by a T1-MPRAGE-sequence (TR = 1900 ms, TE = 2.52 ms, flip angle = 9°, FOV = 250 mm, voxel size = 1 × 1x1 mm^3) as well as a T2 and diffusion-weighted imaging (not analyzed here)

2.4. MRI data analysis

fMRI data were analyzed in SPM8 (Wellcome Trust Centre for Neuroimaging, London, UK) implemented in the Matlab environment. The first five EPI images of each run were discarded before preprocessing. Preprocessing included spatial realignment to the first image, co-registration of the mean functional image to the anatomical MPRAGE image, normalization to the MNI template via segmentation and spatial smoothing with a Gaussian filter with 6 mm FWHM. Subjects with head movements exceeding translations of 3 mm were discarded from the respective analysis. EPI images were temporally high-pass filtered with a cut-off of 128 s. A general linear model (GLM) was then fitted for each subject and task to model the individual BOLD signal changes induced by experimental conditions. Regressors were boxcar functions corresponding in length to the experimental condition (block-type regressors), except for the electro-tactile stimulation in Task 3 and the trials in Task 6, for which we used stick functions (event-related regressors). All analyses were corrected for serial correlated errors by fitting a first-order autoregressive process (AR[1]) to the error term. For Task 6, the contrast images of the three runs of the task were averaged before being entered into group analysis. Random-effects analyses were performed on the group level using single-subject beta or (in Task 6) contrast images of interest in SPM's flexible factorial design.

2.5. Replicability analysis

Replicability of fMRI activations was assessed with the cluster overlap method (Maitra, 2010; Raemaekers et al., 2007) and the Intra Class Correlation (ICC) (Bennett and Miller, 2010; Shrout and Fleiss, 1979).

The spatial overlap was estimated for whole brain-level activation using the Jaccard index (Maitra, 2010). The Jaccard index corresponds to the ratio of commonly activated voxels to the number of voxels activated in only one of the two samples and can thus be interpreted as the percentage of shared significant voxels.

$$Jaccard\ index = \frac{Voverlap}{(V1 + V2 - Voverlap)}$$

$V_{overlap}$: number of voxels active in both samples.
V_1: number of voxels active in sample 1.
V_2: number of voxels active in sample 2.

As preparation for the computation of the Jaccard index, activation maps for our main contrasts of interest were thresholded at $p_{unc}<0.01$ for each sample. This threshold is chosen arbitrarily to ensure adequate number of activated voxels across different tasks and the two samples for comparisons.

In contrast to the Jaccard index the ICC is not dependent on a predefined statistical threshold (Bennett and Miller, 2010). The ICC is commonly used to assess the stability of inter-individual differences in test-retest fMRI study designs (e.g. Moessnang et al., 2016; Plichta et al., 2012; Raemaekers et al., 2007). Additionally, the ICC has been applied to assess the test-retest reliability of fMRI group activations (Plichta et al., 2012; Raemaekers et al., 2007). We here employed the ICC to investigate the replicability of group activations across two independent samples. Turner et al. (2018) used a Pearson correlation for this purpose. The ICC is comparable to a correlation in that it reflects a measure of relatedness between two variables; however, the ICC uses the modelling framework of analysis of variance (ANOVA). In particular, we used the two factors 'voxel' and 'sample' to explain the observed variance in the data, i.e., the variance in *t-scores* between voxels and samples. If the mean square of variance in *t-scores* between voxels (MS_{voxels}) is high compared to the unsystematic error (MS_{error}), then the replicability of group activations can be considered high. To this end, we computed the ICC according to the formula below (Shrout and Fleiss, 1979).

Table 3
Replicability of fMRI tasks.

Contrast	ROI	Jaccard	ICC(2,1)	CI95%
Task 2: Reward sensitivity				
Anticipation				
Gain > Zero (T2C1)	whole brain	0.87	0.77*	[0.48; 0.87]
	R NAcc (R VS)[c]		0.34 (0.65)	[-0.05, 0.71] ([-0.08, 0.88])
Loss > Zero (T2C2)	whole brain	0.77	0.83*	[0.74; 0.88]
	R NAcc (R VS)[c]		0.35 (0.77)	[-0.10; 0.67] ([-0.04; 0.93])
Outcome				
Gain > No Gain (T2C3)	whole brain	0.54	0.70*	[0.68; 0.73]
	B vmPFC		0.53	[-0.09; 0.81]
No loss > Loss (T2C4)	whole brain	0.28	0.47*	[0.03; 0.70]
	L NAcc (L VS)[c]		0.63* (0.56)	[0.50; 0.73] ([-0.05, 0.80])
Task 3: Safety learning and memory				
Conditioning				
CS+s > CS- (T3C1)	whole brain	0.74	0.88*	[0.78; 0.93]
	B dACC		0.72	[-0.03; 0.90]
CS- > CS+s (T3C1rev)	whole brain	0.54	0.88*	[0.78; 0.93]
	B vmPFC		0.87*	[0.84; 0.89]
Extinction				
E-EX > L-EX[a] (T3C2)	whole brain	0.18	0.45*	[0.02; 0.69]
	R AI		0.06	[-0.02; 0.23]
L-EX > E-EX[a] (T3C2rev)	whole brain	0.29	0.45*	[0.02; 0.69]
	R pHPC		0.53*	[0.39; 0.64]
Spontaneous Recovery				
CS+u > CS- (T3C3)	whole brain	0.09	0.26*	[0.18; 0.33]
	R AI		0.13	[-0.07; 0.37]
CS- > CS+u (T3C3rev)	whole brain	0.01	0.26*	[0.18; 0.33]
	B vmPFC		0.18*	[0.09; 0.28]
CS+u > CS+e (T3C4)	whole brain	<0.01	0.04*	[0.02; 0.06]
	R AI		<0.01	[-0.01; 0.01]
Renewal				
CS+s > CS- (T3C5)	whole brain	0.08	0.53*	[0.53; 0.54]
	L AI		0.37*	[0.31; 0.42]
CS- > CS+s (T3C5rev)	whole brain	0.15	0.53*	[0.53; 0.54]
	R OFC		0.09	[-0.06; 0.24]
Task 4: Self-focused volitional reappraisal				
T > NT (T4C1)	whole brain	0.57	0.85*	[0.56; 0.93]
	B rdmPFC		0.73	[-0.06; 0.92]
NT > T (T4C1rev)	whole brain	0.56	0.85*	[0.56; 0.93]
	B vmPFC		0.75*	[0.25; 0.89]
R > NR (T4C2)	whole brain	0.48	0.66*	[0.66; 0.66]
	L dlPFC		0.73*	[0.68; 0.77]
Task 5: Situation-focused volitional reappraisal				
Pos/NR > Neu/NR (T5C1)	whole brain	0.25	0.59*	[0.38; 0.72]
	B vmPFC		0.11	[-0.03; 0.37]
Neg/NR > Neu/NR (T5C2)	whole brain	0.41	0.60*	[0.36; 0.73]
	R amygdala		0.64* (0.58*)[d]	[0.48; 0.76] ([0.41; 071])[d]
R > NR (T5C3)	whole brain	0.56	0.91*	[0.9; 0.92]
	L dlPFC		0.65*	[0.28; 0.81]
Task 6: Emotional interference and motor response inhibition				
Incon > Con Go (T6C1)	whole brain	0.33	0.54	[-0.08 0.80]
	L dlPFC		0.67	[-0.06; 0.90]
Stop > Con Go (T6C2)	whole brain	0.57	0.86*	[0.86; 0.87]
	R IFG – p. tri.		0.57	[-0.04; 0.86]
Neg > Neu (T6C3)	whole brain	0.35	0.84*	[0.83; 0.85]
	R amygdala		0.72	[-0.06; 0.91]
Interaction[b] (T6C4)	whole brain	0.02	0.10	[-0.03; 0.22]
	R amygdala		0.28	[-0.06; 0.64]

Experimental conditions*:* Con Go, congruent go; CS, conditioned stimulus; CS+, CS that is reinforced during conditioning; CS+e, CS that is extinguished during extinction; CS+u, CS that is not extinguished during extinction; CS-, CS that is never reinforced; E-EX, early extinction (first two trials); Incon go, incongruent go; L-EX, late extinction (last two trials); Neg, negative; Neu, neutral; NT, no threat; NR, no reappraisal; Pos, positive; R, reappraisal; T, threat. Anatomical terms: ACC, anterior cingulate cortex; AI, anterior insula; dACC, dorsal anterior cingulate cortex; dlPFC, dorsolateral prefrontal cortex; IFG, p. tri., inferior frontal gyrus - pars triangularis; NAcc, Nucleus acccumbens; pHPC; posterior hippocampus; OFC, orbitofrontal cortex; rdmPFC, rostral dorsomedial prefrontal cortex; vmPFC, ventromedial prefrontal cortex; VS, ventral striatum.

*$p < 0.05$; CI95%, 95% confidence interval; ICC, Intra Class Correlation.

[a] CSs (CS+e, CS-) in early vs. late extinction.

[b] The interaction effect compares the response to negative pictures as compared to neutral ones in congruent go vs. response inhibition trials ((Con Go)>(Incon Go, Stop)/2)$_{Neg}$ > ((Con)>(Incon Go, Stop)/2)$_{Neu}$).

[c] To explore the finding of the unexpected low ICC for the R NAcc ROI, we repeated our replicability analysis using a mask of the ventral striatum (Garrison et al., 2013).

[d] Values after exclusion of single outlier (<mean-3SD).

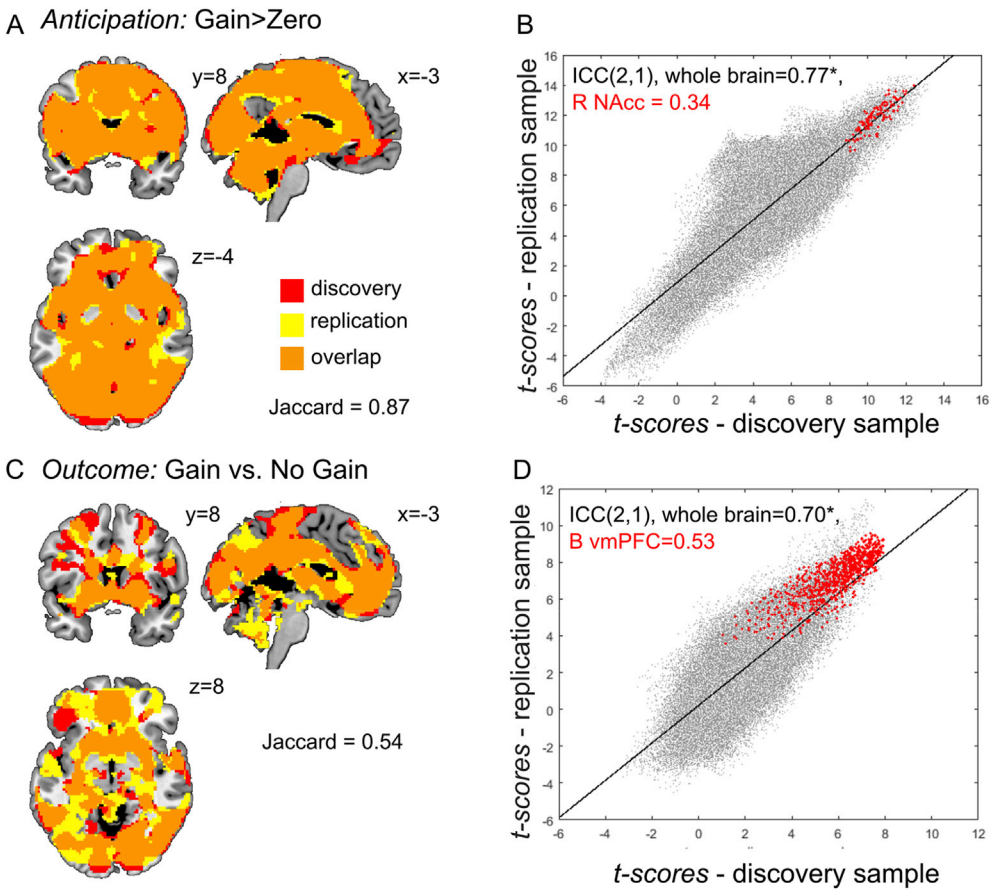

Fig. 2. Task 2; Reward sensitivity (gain anticipation and outcome): fMRI replicability analysis. For all replicability analyses, the spatial overlap (here A, C) as well as the scatter plots for the *t-scores* of the discovery and replication samples (here B, D) are given. All activation maps are thresholded at $p_{unc}<0.01$. Orange voxels in the maps were active in both samples, light blue voxels were deactivated in both samples. The scatter plots are based on all voxels within the brain selected by the SPM group masks from the second-level analyses. Voxels within ROIs are depicted in red if they were active for the contrast and blue if they were active for the reverse contrast. The Intra Class Correlation (ICC(2,1)) indicates the degree of relatedness between the two samples (Shrout and Fleiss, 1979). R NAcc, right nucleus accumbens; B vmPFC, bilateral ventromedial prefrontal cortex.

$$ICC(2,1) = \frac{MSvoxels - MSerror}{MSvoxels + (S-1)MSerror - S(MSsamples - MSerror)/V}$$

MS_{voxels}: mean square of variance in t-scores between voxels.

$MS_{samples}$: mean square of variance in t-scores between sessions or samples.

MS_{error}: error variance.

S: number of samples.

V: number of voxels.

ICC(2,1) models the variance of the sample as random. This means that it expresses the absolute agreement of the fMRI group activations considering group differences in the amplitude of *t-scores*. It can thus be understood as a replicability index that can be generalized to other samples (Shrout and Fleiss, 1979). According to the recommendation by Cicchetti (1994), ICC values < 0.4 indicate poor correspondence, values between 0.4 and 0.59 fair correspondence, values between 0.6 and 0.74 are good and ICC values > 0.75 indicate excellent correspondence.

Replicability analyses were performed with a custom-made Matlab script that is available upon request and an open source Matlab function for computation of the ICC.[4] Before computation of the ICC, voxels were selected using the conjunction of the two brain masks generated by SPM during the second-level analyses of the first and the second sample. This is done to exclude voxels that SPM has discarded from statistical analysis in one of the two samples. ICCs were computed for both the whole brain and for pre-defined regions of interest (ROIs; see task descriptions and Fig. 1).

Two principled differences between the Jaccard index and the ICC are worth noting. First, as a consequence of thresholding, the Jaccard index is specific to the direction of an experimental contrast. Hence, full description of the replicability of the effect of two contrasted regressors requires two Jaccard indices, but only one ICC. Second, the Jaccard index does not take into account the activation amplitude (e.g., *t-scores*) of supra-threshold voxels. It therefore only reflects correspondence of two samples with respect to the spatial distribution, or shape, of an activation. By contrast, the ICC is *t-score*-based and can therefore be seen as indicating correspondence in activation profiles, or landscapes of *t-scores*. If *t-scores* are systematically high in a given area and low in another area, this will increase the ICC. On this basis, one can consider the ICC as a richer source of information on replicability than the Jaccard index. For these reasons, we consider the ICC as the preferable replicability index and base our replicability judgements on this index. Nonetheless, for an open and transparent comparison to previous studies that used measures of spatial overlap (e.g. Maitra, 2010; Moessnang et al., 2016; Plichta et al., 2012; Raemaekers et al., 2007), we also present the Jaccad index for reader's judgement.

To further corroborate how different factors can influence these replicability indices, we performed two additional analyses. Firstly we tested how the Jaccard indices change when we vary our statistical threshold ($p_{FWE}<0.05$, $p_{unc}<0.005$, $p_{unc}<0.01$ and $p_{unc}<0.05$). Then we checked whether the robustness of brain activation in a certain contrast significantly predicts the respective ICC. Due to the small number of datapoints, we used both ordinary correlation and robust regression analysis implemented in Matlab.

3. Results

3.1. Replicability of fMRI tasks

Results for the behavioral and physiological data are given in suppl. materials S8. Table 3 gives an overview on the replicability measures

[4] Salarian, A. (2008). Function for computation of the Intraclass Correlation Coefficient (ICC). Retrieved from https://de.mathworks.com/matlabcentral/fileexchange/22099-intraclass-correlation-coefficient–icc-?

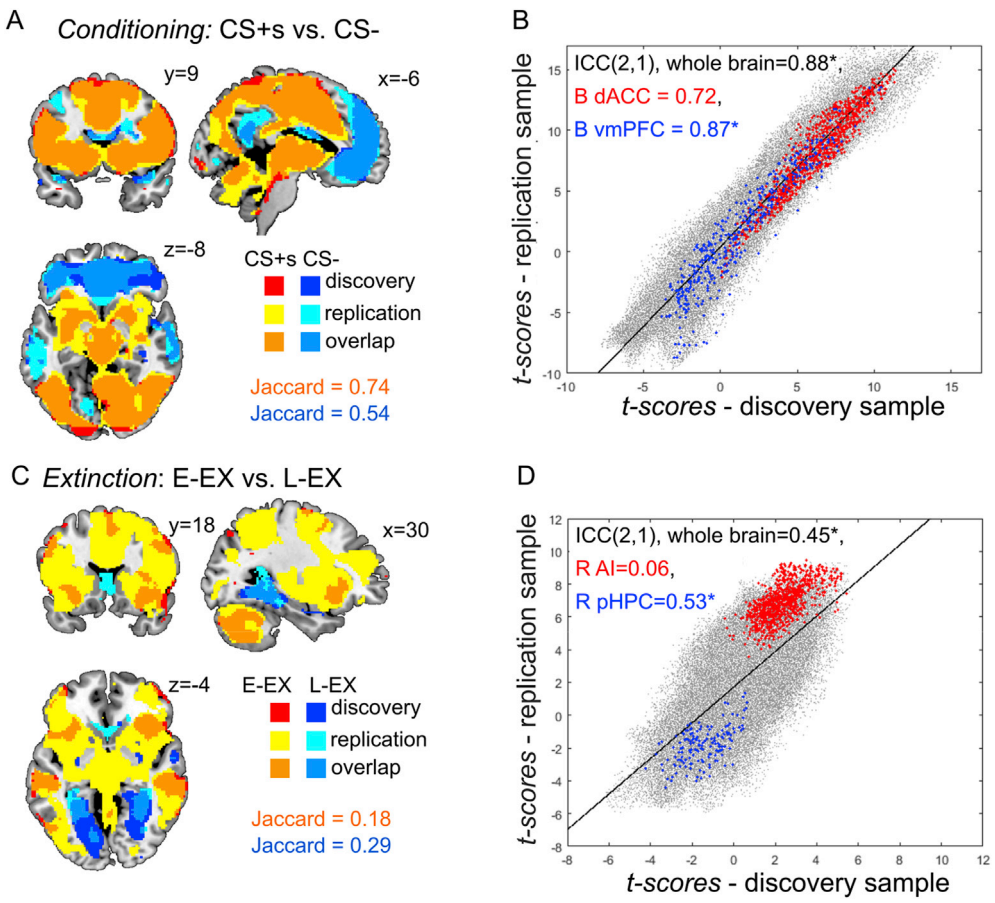

Fig. 3. Task 3: Safety learning and memory (conditioning and extinction): fMRI replicability analysis. CS+, CS that is reinforced during conditioning; CS-, CS that is never reinforced; AI, anterior insula; dACC, dorsal anterior cingulate cortex; pHPC, posterior hippocampus; vmPFC, ventromedial prefrontal cortex.

(Jaccard index, ICC(2,1)) for the different fMRI tasks. Both indices remained stable across tasks when we controlled for the group differences in age and negative life events between the two samples by adding these two factors as a covariate in the fMRI second-level models (see suppl. materials S13 Table S2).

3.1.1. Task 2: Reward sensitivity

In the anticipation phase, the Jaccard index indicated a high spatial overlap of supra-threshold voxels between the two samples both for gain (T2C1) and loss anticipation (T2C2). Additionally, the whole brain ICCs, expressing correspondence in *t-score* activation profiles, were excellent (see Table 3). By contrast, ICCs in the R NAcc ROI were poor for both contrasts. To explore this unexpected finding we repeated our analysis with a ventral striatum ROI which resulted in good ICCs for the anticipation of gain and loss (see Table 3). In the outcome phase, Jaccard indices for both gain (T2C3) and no loss outcomes (T2C4) were weaker (fair to good), as were the ICCs for the whole brain. For gain outcome, the B vmPFC ROI showed fair replicability; for no loss outcome, the ICC in the L NAcc ROI was good. Fig. 2 shows the spatial overlap of activations and a scatter plot for the *t-scores* of the two samples for the main contrasts of interest gain anticipation and gain outcome. See suppl. materials S9 for the figure on loss anticipation and no loss outcome Overall, results indicate that the task is well replicable at the whole brain but less so at ROI level.

3.1.2. Task 3: Safety learning and memory

In conditioning, spatial activation patterns from both the CS+s > CS- (T3C1) and the reverse CS- > CS+s (T3C1rev) contrast were highly replicable (see Table 3, Fig. 3 A). The ICC for the whole brain was excellent. ICCs for the ROIs (B dACC, B vmPFC) were good to excellent (Fig. 3 B). In extinction, spatial pattern overlap between samples was poor for both contrasts (Table 3, Fig. 3 C). The whole brain ICC was fair (see Fig. 3 D). The ICC for the R AI ROI in the E-EX > L-EX contrast (T3C2) was poor, while the ICC for the R pHPC ROI in the L-EX > E-EX (T3C2rev) contrast was fair. In spontaneous recovery, replicability was poor across contrasts (see Table 3 and suppl. materials S10). In line with former work we tested CS+u > CS- (T3C3), the reverse contrast CS- > CS+u (T3C3rev) and CS+u > CS+e (T3C4) (Kampa et al., 2018). In renewal, spatial activation patterns (Jaccard index) were also poorly replicated (see Table 3 and suppl. materials S10), however, the whole brain ICC was fair for both the CS+s > CS- (T3C5) and the CS- > CS+s (T3C5 rev) contrast. ROI ICCs were poor. Overall, the applied conditioning task appears highly replicable, extinction and renewal are only moderately replicable, while spontaneous recovery is not.

3.1.3. Task 4: Self-focused volitional reappraisal

In the threat (T4C1) and the safety contrast (T4C1rev), activation patterns showed good spatial overlap, the whole brain ICC was excellent and ROI ICCs were good to excellent (see Table 3, Fig. 4 A and B). In the reappraisal contrast (T4C2), spatial overlap of activation patterns was somewhat smaller and all ICCs were in the good range (Fig. 4 C and D). Overall, the threat manipulation is highly replicable, the reappraisal manipulation is well replicable.

3.1.4. Task 5: Situation-focused volitional reappraisal

In the positive (T5C1) and negative (T5C2) picture viewing contrasts, activation patterns showed moderate spatial overlap, while both whole brain and ROI ICCs were good, with exception of the poor ICC for the B vmPFC ROI in positive picture viewing (Table 3, Fig. 5 A–D). Exclusion of the outlier for the R amygdala ROI in negative picture viewing, slightly

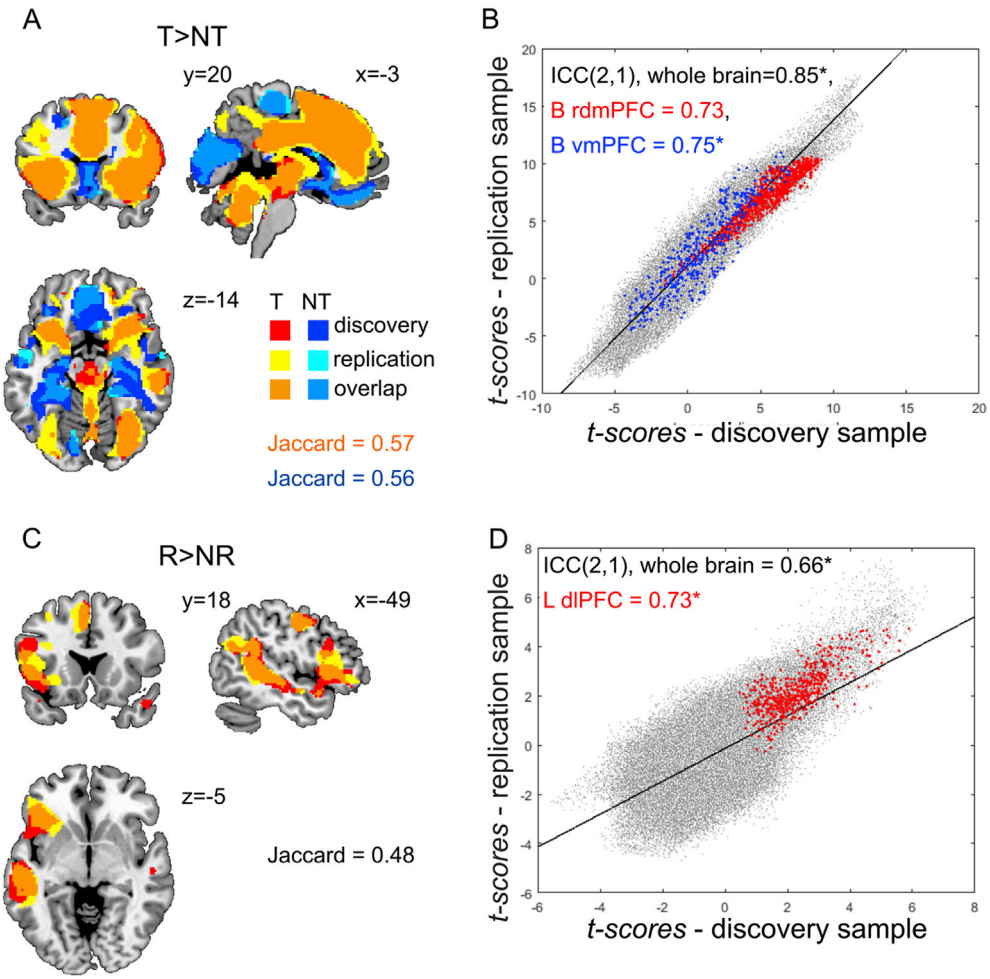

Fig. 4. Task 4: Self-focused volitional reappraisal: fMRI replicability analysis. dlPFC, dorsolateral prefrontal cortex; rdmPFC, rostral dorsomedial PFC; vmPFC, ventromedial PFC; NR, no reappraisal; NT, no threat; R, reappraisal; T,threat.

reduced the ICC (Fig. 5 D). The reappraisal contrast (T5C3) showed considerably better spatial overlap and an excellent whole brain ICC. The ICC for the L dlPFC ROI was good (Fig. 5 E, F). Overall, the picture viewing contrasts are well and the reappraisal contrast is very well replicable.

3.1.5. *Task 6: Emotional interference and motor response inhibition*

Both the spatial interference inhibition (Incon > Con Go, T6C1) and the action cancellation (Stop > Con Go, T6C2) contrasts, showed good spatial overlap (Table 3, Fig. 6 A–D). Whole brain ICCs were fair to excellent. ROI ICCs (L dlPFC, R IFG, p. tri.) were fair to good. In the negative primes contrast (T6C3) (Fig. 6 E, F), spatial overlap was good and the whole brain ICC was excellent, the ICC for the R amygdala ROI was good. In contrast to the main effects all ICCs for the interaction contrast (T6C4) were poor. The ICC for the R amygdala ROI was slightly higher than the whole brain ICC implicating that replicability was stronger for the ROI. See suppl. materials S12 for a figure of the interaction effect. Except for the interaction contrast the task can be considered highly replicable.

3.2. *Role of activation threshold for replicability of spatial activation overlap*

Because calculation of the Jaccard index requires activation thresholding, we asked if spatial overlap changes with different thresholds ($p_{FWE}<0.05$, $p_{unc}<0.005$, $p_{unc}<0.05$ and $p_{unc}<0.01$). Fig. 7 shows that results in our samples are similar irrespective of the chosen threshold. When we use a more stringent statistical threshold ($p_{FWE}<0.05$), significantly lower Jaccard indices were observed. However the overall pattern of Jaccard indices under $p_{FWE}<0.05$ remained highly similar to those obtained from uncorrected thresholds.

3.3. *Role of task activation strength*

Turner et al. (2018) had suggested that tasks with stronger overall activation show better replicability. We therefore correlated the number of supra-threshold voxels at $p_{unc}<0.01$ in each contrast from the discovery sample with the contrasts' whole brain ICCs. The correlation was significantly positive ($r = 0.4965$, $p = 0.059$; using robust regression $r = 0.5103$, $p < 0.001$; see Fig. 8).

4. Discussion

In this manuscript we tested the replicability of fMRI activations in a test battery assessing various forms of stress reactivity and regulation and associated constructs in two independent cohorts of healthy young adults. The two samples differ only in terms of age and history of negative life events due to the study's inclusion criteria, but controlled for other aspects of psychosocial functioning. We used the Jaccard index and the ICC to quantify replicability of group-based brain activation patterns

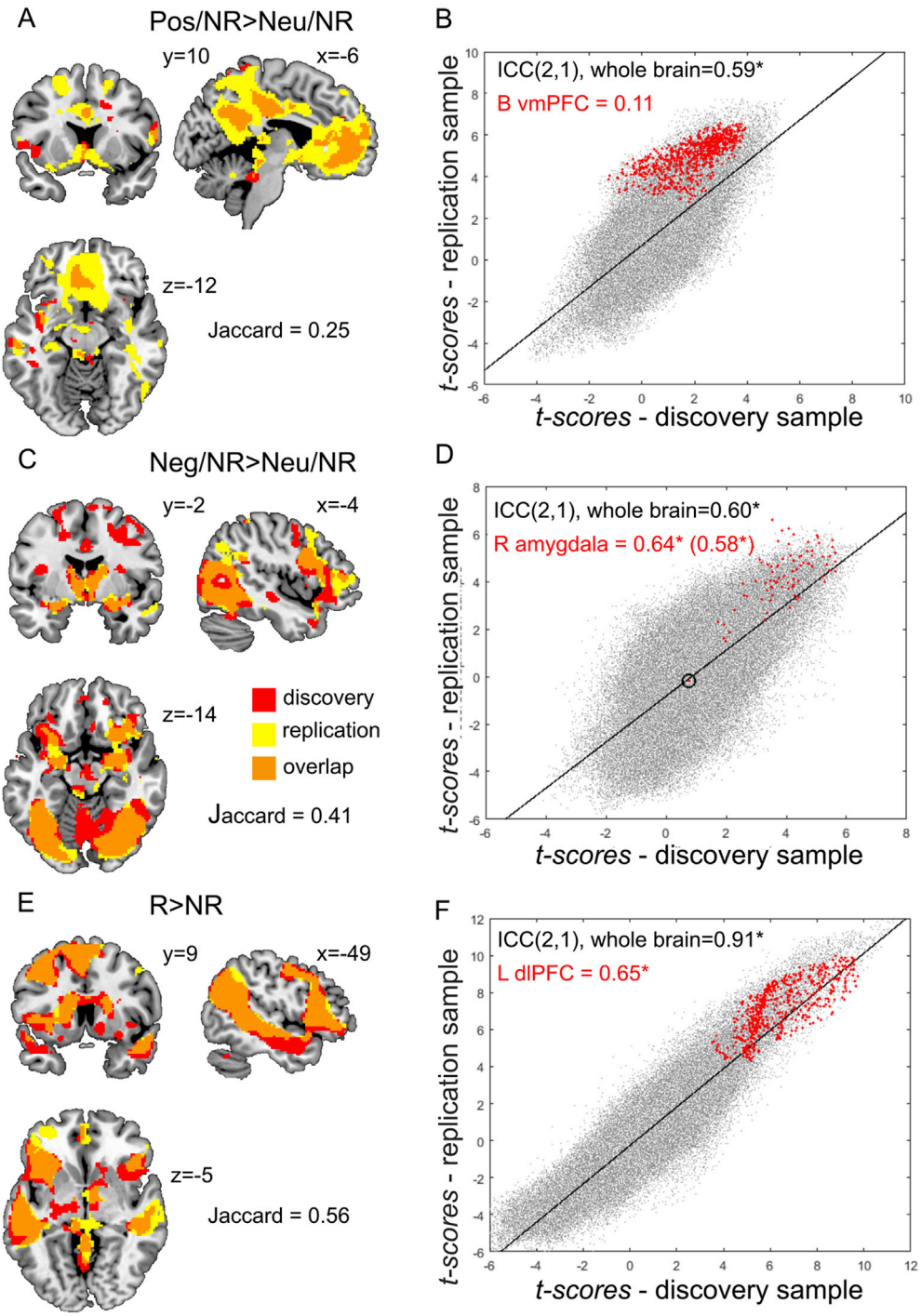

Fig. 5. Task 5: Situation-focused reappraisal: fMRI replicability analysis. D) The ROI ICC after exclusion of the indicated outlier (<mean-3SD) is given in brackets. Neg, negative; Neu, neutral; NR, no reappraisal; Pos, positive; R, reappraisal; dlPFC, dorso lateral prefrontal cortex; vmPFC; ventromedial PFC.

across different tasks.

4.1. Replicability of the neuroimaging test battery

Based on ICC results, satisfactory replicability at the whole brain level could be shown for the majority of task contrasts. This included all tested contrasts in the reward sensitivity task (Task 2), the conditioning and – to a lesser extent – the extinction and renewal contrasts in the safety learning and memory task (Task 3), the threat and reappraisal contrasts in the self-focused volitional reappraisal task (Task 4), the picture viewing and reappraisal contrasts in the situation-focused volitional reappraisal task (Task 5), and the two motor response inhibition and – to a lesser extent – the negative picture primes contrast in the interference inhibition task (Task 6). By contrast, the spontaneous recovery contrast in Task 3 and the emotion by motor response inhibition interaction in Task 6 did not replicate well. One explanation for the low replicability of the spontaneous recovery contrast could be a high rate of inter-individual differences (Lonsdorf and Merz, 2017). Some subjects might retrieve their extinction memory, while others might show return of fear, leading to a huge variance in activation patterns and a low group effect size. Concerning the low reproducibility of the interaction effect, one needs to consider that the standard error of an interaction term is much higher

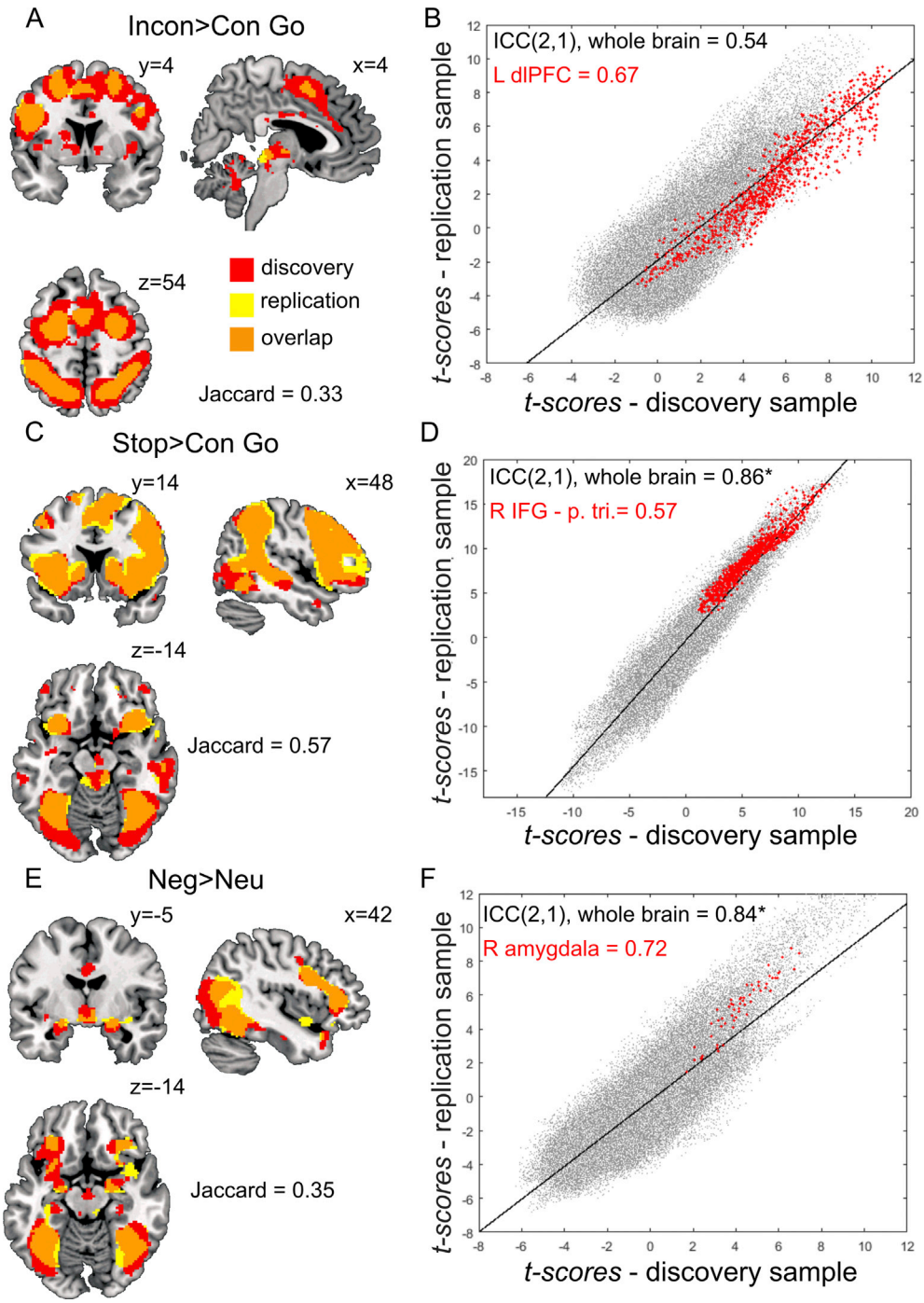

Fig. 6. Task 6: Emotional interference and response inhibition: fMRI replicability analysis. Con go, congruent go; Incon Go, incongruent go; Neg, negative; Neu, neutral, dlPFC, dorso lateral prefrontal cortex; Incon Go, incongruent go; IFG - p. tri.; inferior frontal gyrus - pars triangularis. Results from the interaction contrast is presented in the suppl. materials S12.

than for a main effect because it depends on the measurement error of both interacting variables (Frazier et al., 2004; McClelland and Judd, 1993). Replication of an interaction is thus more difficult than replication of a main effect and requires higher sample sizes.

Our current sample sizes (discovery: N = 54, replication: N = 103) are well over the conventional standard of reported studies (median sample size of fMRI studies in 2015 = 28.5; Poldrack et al., 2017). As such why particular contrasts are not replicable is worth further investigation. With a closer look at the findings obtained from the discovery sample, these non-replicable contrasts stood out by showing by far the weakest effect sizes both in terms of strength and extend of activation (Kampa et al., 2018). This prompted us to examine the relationship between the robustness of brain activations in the discovery sample and the subsequent replicability in the replication sample. Using the number of supra-threshold voxels as a proxy for robustness of activations and ICC as a proxy for replicability, we observed a significant positive correlation between the two. In other words, provided adequate sample size, the more robust the activation a task induces at a group level, the higher the chance that the same pattern of brain activations can be observed in an independent sample. Considering the heterogeneity of tasks in the test

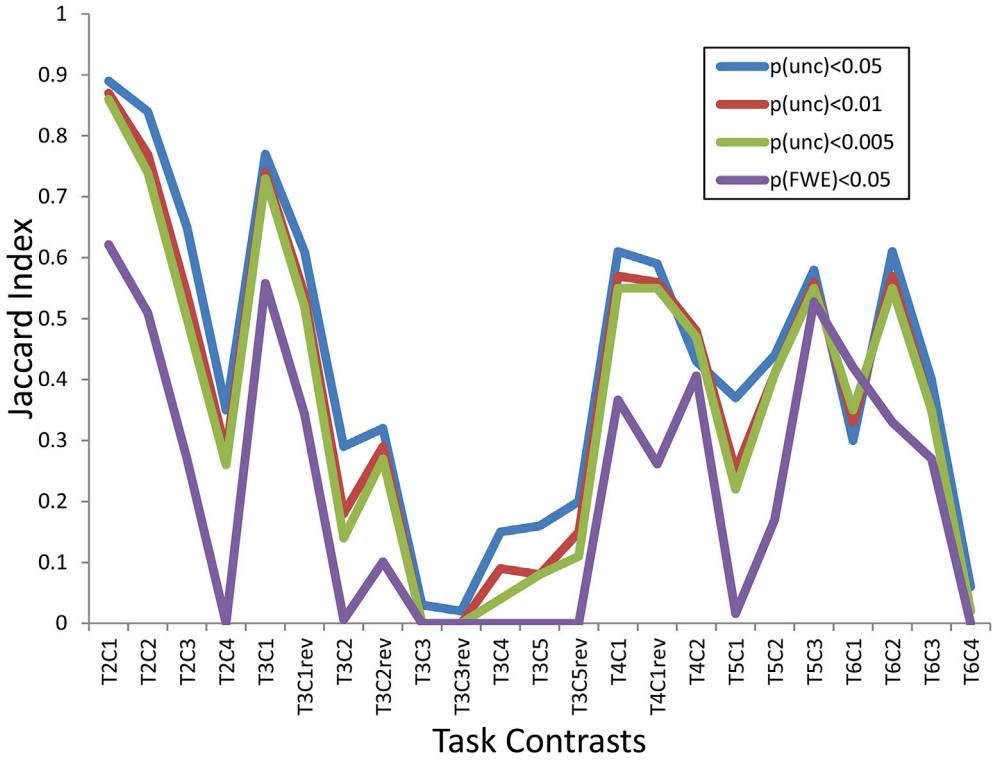

Fig. 7. Jaccard indices under different statistical thresholds across all task contrasts. For contrast codes (C) under a certain task (T) please refer to Table 3.

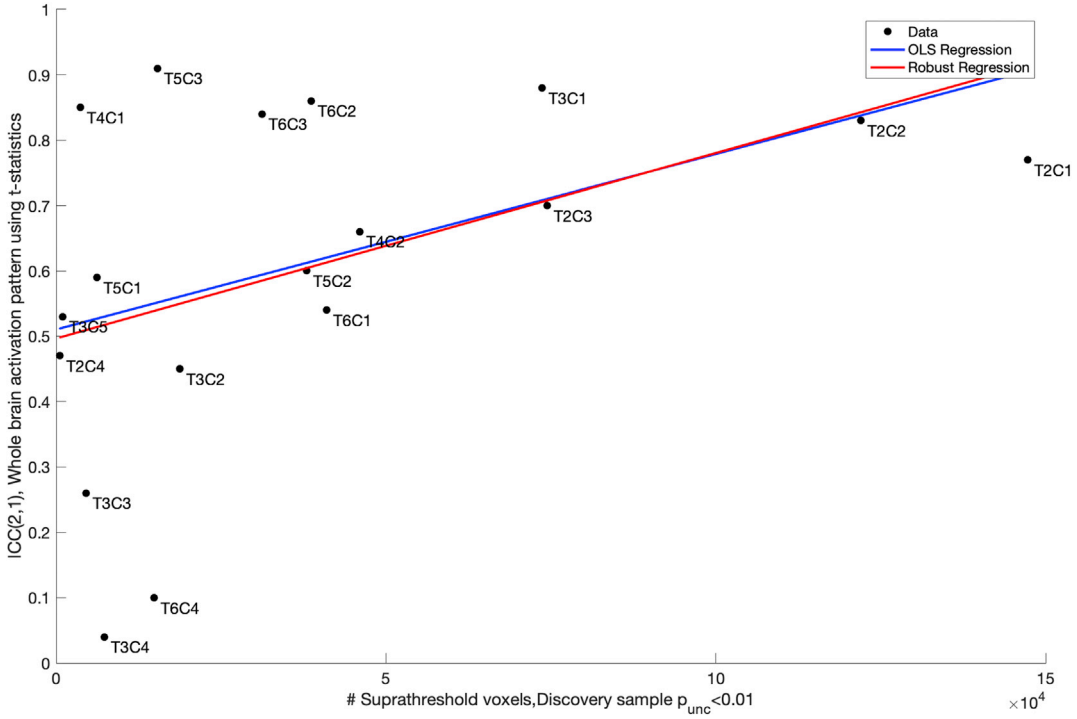

Fig. 8. Significant positive correlation between whole brain ICC and robustness of task activations (as indicated by number of supra-threshold voxels p = 0.01 uncorrected).

battery, our observation on the variability of replicability in different task contrasts falls well in line with the recent findings reported by Turner et al. (2018). They observed that even with a sample size of 100, replicability in terms of whole brain activation pattern still varies substantially between different tasks.

4.2. *Replicability in ROIs*

Apart from using whole brain activation patterns for statistical inferences, researchers with *a priori* hypotheses may also restrict their search space to a few regions-of-interest. In the current analysis we test this ROI-based replicability in our task contrasts. Overall, when we

restrict our ICC calculation to the most robust ROI observed in the discovery sample, a slightly lower ICC than the whole brain pattern ICC is obtained which may seem counterintuitive. This can be expected as the size of the ROIs (range: 110–925 voxels) are approximately only a thousandth of the whole brain volume; which can render the agreement between two samples lower due to the use of fewer data points. Another aspect contributing to a lower ICC is that the variance for voxels within a ROI is lower since it includes mostly supra-threshold *t-scores* lying either in the upper or lower part of the distribution. Finally the ROI ICC seems to be more sensitive to differences in the mean amplitude of *t-scores*. This means, that even though the correlation between the two samples might be high (relative agreement), the ICC(2,1) used here (absolute agreement) could still be low due to differences in the amplitude of the two samples (e.g. R NAcc in gain and loss anticipation). Nonetheless, a closer scrutiny (Figs. 2–6, right panels) revealed that the relationship of the two samples within the ROIs often conforms to the relationship observed in the whole brain patterns. When judging the replicability for one ROI one should of course not only focus on the ICC but also inspect if the ROI exceeds the statistical threshold, thus indeed replicating the effect. Concluding from our present results the ICC(2,1) which is used here seems better suited to judge the replicability on whole brain than on ROI level.

4.3. Group replicability and inter-individual differences

The ultimate goal of the current test battery is to gain a mechanistic understanding on the neurocognitive processes underlying stress reactivity and regulation, which in the long run predicts an individual's psychological resilience against adversities. For the test battery to be used as a valid tool for such purpose, the tasks should be reliable within subject, replicable across samples, and sensitive enough to capture individual differences. Due to resources constrains we can only test the replicability of task contrasts in the current study but not its reliability.

An observation that has been frequently made, is that the reliability (Caceres et al., 2009; Lois et al., 2018; Moessnang et al., 2016; Nord et al., 2017; Plichta et al., 2012; Raemaekers et al., 2007) and also the replicability of fMRI group activations is high (Turner et al., 2018). Yet reliability for group activations does not equal to stability of inter-individual differences within the group, which is reportedly lower (e.g. Caceres et al., 2009; Infantolino et al., 2018; Lois et al., 2018; Nord et al., 2017). Interestingly, Caceres et al. (2009) revealed that voxels with a high effect size on group level were also more reliable for assessing inter-individual differences than voxels with a low effect size. Thus identifying regions with robust group activation could indirectly point to regions that are also reliable for inter-individual differences. In contradiction to the assumption of a positive relationship between replicability and reliability stands the repeated observation that tasks with robust activation often show poor reliability for inter-individual differences (Hedge et al., 2018; Infantolino et al., 2018). An explanation could be that the relationship of robust activation and reliability, is moderated by factors, like characteristics of the task (Lois et al., 2018). In cases where low inter-individual variation is achieved by e.g. strong experimental control low reliability can actually be associated with good replicability (De Schryver, Hughes, Rosseel and De Houwer, 2016). In conclusion, the test-retest reliability on individual level and thus the stability of inter-individual differences should be tested in the longitudinal follow-up of the MARP subjects to guarantee not only replicability of group activations but also reliability for inter individual differences.

4.4. Limitations

As we have already stated in the introduction, we did not have the chance to perform a test-retest reliability analyses since samples were only tested once. Another limitation is that samples were significantly different from each other in age, number of life events, employment status and in the coping style self-distraction which is mainly due to the use of different inclusion criteria and could have reduced replicability. Nonetheless all subjects in MARP are screened for psychiatric problems and they all have normal psychiatric functioning at the point of baseline testing. Another issue is related to the Jaccard index. We are not aware of any conventional standards for its interpretation. As we have shown under 3.2 it is influenced by the chosen statistical threshold. As a consequence the interpretation of our replicability analysis is mainly based on the ICC values.

The high temporal resolution resulting from the use of multiband EPI sequence increases the risk for false positives by underestimating the temporal autocorrelation during our conventional first level SPM analysis (Bollmann et al., 2018; Corbin et al., 2018; Demetriou et al., 2018; McDowell and Carmichael, 2019). In order to better model the temporal autocorrelation in multiband EPI sequences an improved algorithm called the FAST model is being introduced. Nonetheless it is still inconclusive how the use of FAST algorithm will influence group level results (Bollmann et al., 2018; Corbin et al., 2018). Since our initial analyses on the discovery sample were performed in SPM8 we had used AR(1) model for temporal autocorrelation, before the existence of the FAST algorithm, our current calculation of Jaccard and ICC are all based on the AR(1) model at the first level analyses. Considering that the modeling of temporal autocorrelation might be a potential confound, we repeated our first level analyses using the FAST algorithm in SPM12 instead of AR(1) in SPM8. Results are given in the suppl. materials S14. Overall results of the replicability analyses of the two analyses converge well.McDowell and Carmichael, 2019

Although we have not tested this explicitly in our current analysis, we want to emphasize that the use of different preprocessing pipeline might also impact on the replicability of task activities. In our current analysis, we have used a standard preprocessing pipeline that reflects the most common practice in the field. More specifically, we used an epi-to-mprage normalization method based on unified segmentation that was implemented since SPM5. Recently it has been suggested (Calhoun et al., 2017) that epi-to-epi normalization might improve inter-subject alignment, thereby increasing statistical power. We expect that any improvements in terms of statistical power brought by different preprocessing pipeline will further improve the replicability of task activations.

5. Conclusion

The recent replicability crisis, emerging from the frequent failure to reproduce results of other working groups or in other samples, led to a decrease of reputation of psychology and neuroimaging research (e.g. Evans, 2017; Open Science Collaboration, 2015). In the present manuscript we demonstrated good to excellent replicability across different task domains. With respect to the recent replicability crisis this strengthens the credibility of findings concerning task group activations. By using two distinct samples for discovery and replication that even showed significant differences in demographic variables, we increased the generalizability of our results. As we demonstrated that tasks with stronger activation in the discovery sample showed better replicability and vice versa, one could conclude that the need for replication is even higher if group activations are only weak. It is especially important to use robust tasks when investigating inter-individual differences or endophenotypes of mental disorders to reduce the amount of error variance. In future research the presented test battery will be applied for the investigation of inter-individual differences related to resilience. Highly replicable group activations can guide the selection of ROIs for the analysis of inter-individual differences since voxels are more reliable if they show robust group activations (Caceres et al., 2009).

Acknowledgement

We wish to thank Petra Seyfarth and Manuela Götz for their engagement in subject recruitment, data acquisition and study

organization. Further help was provided by Thomas Bauermann, Hanno Burger, Samira Christmann, Haakon Engen, Sarah Mohr, Victor Saase, Alexander Schüler, Goran Vucurevic, Merle Wachendörfer and Vanessa Zörrer.

This work was funded by Stiftung Rheinland-Pfalz für Innovation [MARP program, grant number 961-386261/1080] and the Ministry of Science of the state of Rhineland-Palatinate [DRZ program]. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777084 (Dyna-MORE project)..

The authors of this paper do not have any commercial associations that might pose a conflict of interest in connection with this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.116223.

References

Adler, N.E., Epel, E.S., Castellazzo, G., Ickovics, J.R., 2000. Relationship of subjective and objective social status with psychological and physiological functioning: preliminary data in healthy, White women. Health Psychol. 19, 586–592. https://doi.org/10.1037/0278-6133.19.6.586.

Baker, M., 2016. 1,500 scientists lift the lid on reproducibility. Nature 533, 452–454. https://doi.org/10.1038/533452a.

Begley, C.G., Ellis, L.M., 2012. Raise standards for preclinical cancer research: drug development. Nature 483, 531–533. https://doi.org/10.1038/483531a.

Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Ann. N. Y. Acad. Sci. 1191, 133–155. https://doi.org/10.1111/j.1749-6632.2010.05446.x.

Bollmann, S., Puckett, A.M., Cunnington, R., Barth, M., 2018. Serial correlations in single-subject fMRI with sub-second TR. NeuroImage 166, 152–166. https://doi.org/10.1016/j.neuroimage.2017.10.043.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–376. https://doi.org/10.1038/nrn3475.

Caceres, A., Hall, D.L., Zelaya, F.O., Williams, S.C.R., Mehta, M.A., 2009. Measuring fMRI reliability with the intra-class correlation coefficient. NeuroImage 45, 758–768.

Calhoun, V.D., Wager, T.D., Krishnan, A., Rosch, K.S., Seymour, K.E., Nebel, M.B., Mostofsky, S.H., Nyalakanai, P., Kiehl, K., 2017. The impact of T1 versus EPI spatial normalization templates for fMRI data analyses. Hum. Brain Mapp. 38, 5331–5342.

Carver, C.S., 1997. You want to measure coping but your protocol's too long: consider the Brief COPE. I. Int. J. Behav. Med. 4, 92–100.

Caspi, A., Moffit, T.E., Thorton, A., 1996. The Life History Calender: a research and clinical assessment method for collecting retrospective event-history data. Int. J. Methods Psychiatr. Res. 6, 101–114.

Chmitorz, A., Wenzel, M., Stieglitz, R.-D., Kunzler, A., Bagusat, C., Helmreich, I., Gerlicher, A., Kampa, M., Kubiak, T., Kalisch, R., Lieb, K., Tüscher, O., 2018. Population-based validation of a German version of the brief resilience scale. PLoS One 13, e0192761. https://doi.org/10.1371/journal.pone.0192761.

Cicchetti, D.V., 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol. Assess. 6, 284–290. https://doi.org/10.1037/1040-3590.6.4.284.

Corbin, N., Todd, N., Friston, K.J., Callaghan, M.F., 2018. Accurate modeling of temporal correlations in rapidly sampled fMRI time series. Hum. Brain Mapp. 39, 3884–3897. https://doi.org/10.1002/hbm.24218.

Dalgard, O.S., Bjork, S., Tambs, K., 1995. Social support, negative life events and mental health. Br. J. Psychiatry 166, 29–34.

Demetriou, L., Kowalczyk, O.S., Tyson, G., Bello, T., Newbould, R.D., Wall, M.B., 2018. A comprehensive evaluation of increasing temporal resolution with multiband-accelerated protocols and effects on statistical outcome measures in fMRI. NeuroImage 176, 404–416. https://doi.org/10.1016/j.neuroimage.2018.05.011.

De Schryver, M., Hughes, S., Rosseel, Y., De Houwer, J., 2016. Unreliable yet still replicable: a comment on LeBel and Paunonen (2011). Front. Psychol. 6. https://doi.org/10.3389/fpsyg.2015.02039.

Euteneuer, F., Süßenbach, P., Schäfer, S.J., Rief, W., 2014. Subjektiver sozialer Status. MacArthur-Skalen zur Erfassung des wahrgenommenen sozialen Status im sozialen Umfeld (SSS-U) und in Deutschland (SSS-D). Verhaltenstherapie 25, 229–232. https://doi.org/10.1159/000371558.

Evans, S., 2017. What has replication ever done for us? Insights from neuroimaging of speech perception. Front. Hum. Neurosci. 11. https://doi.org/10.3389/fnhum.2017.00041.

Feinberg, D.A., Moeller, S., Smith, S.M., Auerbach, E., Ramanna, S., Glasser, M.F., Miller, K.L., Ugurbil, K., Yacoub, E., 2010. Multiplexed Echo Planar Imaging for Sub-Second Whole Brain FMRI and Fast Diffusion Imaging. PLoS ONE. https://doi.org/10.1371/journal.pone.0015710.

Frazier, P.A., Tix, A.P., Barron, K.E., 2004. Testing moderator and mediator effects in counseling psychology research. J. Couns. Psychol. 51, 115–134. https://doi.org/10.1037/0022-0167.51.1.115.

Garnefski, N., Kraaij, V., Spinhoven, P., 2001. Negative life events, cognitive emotion regulation and depression. Personal. Individ. Differ. 30, 1311–1327.

Garrison, J., Erdeniz, B., Done, J., 2013. Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. Neurosci. Biobehav. Rev. 37, 1297–1310. https://doi.org/10.1016/j.neubiorev.2013.03.023.

Glaesmer, H., Hoyer, J., Klotsche, J., Herzberg, P.Y., 2008. Die deutsche Version des Life-Orientation-Tests (LOT-R) zum dispositionellen Optimismus und Pessimismus. Z. für Gesundheitspsychol. 16, 26–31. https://doi.org/10.1026/0943-8149.16.1.26.

Goldberg, D.P., Hillier, V.F., 1979. A scaled version of the general health questionnaire. Psychol. Med. 9, 139–145.

Haller, S.P., Kircanski, K., Stoddard, J., White, L.K., Chen, G., Sharif-Askary, B., Zhang, S., Towbin, K.E., Pine, D.S., Leibenluft, E., Brotman, M.A., 2018. Reliability of neural activation and connectivity during implicit face emotion processing in youth. Dev. Cogn. Neurosci. 31, 67–73. https://doi.org/10.1016/j.dcn.2018.03.010.

Hedge, C., Powell, G., Sumner, P., 2018. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. Behav. Res. Methods 50, 1166–1186. https://doi.org/10.3758/s13428-017-0935-1.

Infantolino, Z.P., Luking, K.R., Sauder, C.L., Curtin, J.J., Hajcak, G., 2018. Robust is not necessarily reliable: from within-subjects fMRI contrasts to between-subjects comparisons. NeuroImage 173, 146–152. https://doi.org/10.1016/j.neuroimage.2018.02.024.

Kalisch, R., 2006. Context-dependent human extinction memory is mediated by a ventromedial prefrontal and hippocampal network. J. Neurosci. 26, 9503–9511. https://doi.org/10.1523/JNEUROSCI.2021-06.2006.

Kampa, M., Schick, A., Yuen, K., Sebastian, A., Chmitorz, A., Saase, V., Wessa, M., Tüscher, O., Kalisch, R., 2018. A Combined Behavioral and Neuroimaging Battery to Test Positive Appraisal Style Theory of Resilience in Longitudinal Studies. https://doi.org/10.1101/470435.

Kanske, P., Heissler, J., Schonfelder, S., Bongers, A., Wessa, M., 2011. How to regulate emotion? Neural networks for reappraisal and distraction. Cerebr. Cortex 21, 1379–1388. https://doi.org/10.1093/cercor/bhq216.

Kemper, C.J., Ziegler, M., Taylor, S., 2009. Überprüfung der psychometrischen Qualität der deutschen Version des Angstsensitivitätsindex-3. Diagnostica 55, 223–233. https://doi.org/10.1026/0012-1924.55.4.223.

Klaiberg, A., Schumacher, J., Brähler, E., 2004. General Health Questionnaire 28 (GHQ-28): teststatistische Überprüfung einer deutschen Version in einer bevölkerungsrepräsentativen stichprobe. Z. Klin. Psychol. Psychiatr. Psychother. 52, 31–42.

Knoll, N., Rieckmann, N., Schwarzerq, R., 2005. Coping as a mediator between personality and stress outcomes: a longitudinal study with cataract surgery patients. Eur. J. Personal. 19, 229–247.

Knutson, B., Adams, C.M., Fong, G.W., Hommer, D., 2001. Anticipation of increasing monetary reward selectively recruits nucleus accumbens. J. Neurosci. 21.

Kristo, G., Rutten, G.-J., Raemaekers, M., de Gelder, B., Rombouts, S.A.R.B., Ramsey, N.F., 2014. Task and task-free FMRI reproducibility comparison for motor network identification: task and task-free fMRI reproducibility comparison for motor network identification. Hum. Brain Mapp. 35, 340–352. https://doi.org/10.1002/hbm.22180.

LeBel, E.P., Paunonen, S.V., 2011. Sexy but often unreliable: the impact of unreliability on the replicability of experimental findings with implicit measures. Personal. Soc. Psychol. Bull. 37, 570–583. https://doi.org/10.1177/0146167211400619.

Loch, N., Hiller, W., Witthöft, M., 2011. Der Cognitive Emotion Regulation Questionnaire (CERQ): Erste teststatistische Überprüfung einer deutschen adaption. Z. Klin. Psychol. Psychother. 40, 94–106. https://doi.org/10.1026/1616-3443/a000079.

Lois, G., Kirsch, P., Sandner, M., Plichta, M.M., Wessa, M., 2018. Experimental and methodological factors affecting test-retest reliability of amygdala BOLD responses. Psychophysiology e13220. https://doi.org/10.1111/psyp.13220.

Lonsdorf, T.B., Merz, C.J., 2017. More than just noise: inter-individual differences in fear acquisition, extinction and return of fear in humans - biological, experiential, temperamental factors, and methodological pitfalls. Neurosci. Biobehav. Rev. 80, 703–728. https://doi.org/10.1016/j.neubiorev.2017.07.007.

Maitra, R., 2010. A re-defined and generalized percent-overlap-of-activation measure for studies of fMRI reproducibility and its use in identifying outlier activation maps. NeuroImage 50, 124–135. https://doi.org/10.1016/j.neuroimage.2009.11.070.

McClelland, G.H., Judd, C.M., 1993. Statistical difficulties of detecting interactions and moderator effects. Psychol. Bull. 114, 376–390. https://doi.org/10.1037/0033-2909.114.2.376.

McDowell, A.R., Carmichael, D.W., 2019. Optimal repetition time reduction for single subject event-related functional magnetic resonance imaging. Magnetic Resonance in Medicine 81, 1890–1897. https://doi.org/10.1002/mrm.27498.

Milad, M.R., Wright, C.I., Orr, S.P., Pitman, R.K., Quirk, G.J., Rauch, S.L., 2007. Recall of fear extinction in humans activates the ventromedial prefrontal cortex and Hippocampus in concert. Biol. Psychiatry 62, 446–454. https://doi.org/10.1016/j.biopsych.2006.10.011.

Moessnang, C., Schäfer, A., Bilek, E., Roux, P., Otto, K., Baumeister, S., Hohmann, S., Poustka, L., Brandeis, D., Banaschewski, T., Meyer-Lindenberg, A., Tost, H., 2016. Specificity, reliability and sensitivity of social brain responses during spontaneous mentalizing. Soc. Cogn. Affect. Neurosci. 11, 1687–1697. https://doi.org/10.1093/scan/nsw098.

Munafò, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J.J., Ioannidis, J.P.A., 2017. A manifesto for reproducible science. Nat. Human Behav. 1, 0021. https://doi.org/10.1038/s41562-016-0021.

Nettekoven, C., Reck, N., Goldbrunner, R., Grefkes, C., Weiß Lucas, C., 2018. Short- and long-term reliability of language fMRI. NeuroImage 176, 215–225. https://doi.org/10.1016/j.neuroimage.2018.04.050.

Nord, C.L., Gray, A., Charpentier, C.J., Robinson, O.J., Roiser, J.P., 2017. Unreliability of putative fMRI biomarkers during emotional face processing. NeuroImage 156, 119–127. https://doi.org/10.1016/j.neuroimage.2017.05.024.

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349 aac4716–aac4716. https://doi.org/10.1126/science.aac4716.

Paret, C., Brenninkmeyer, J., Meyer, B., Yuen, K.S.L., Gartmann, N., Mechias, M.-L., Kalisch, R., 2011. A test for the implementation-maintenance model of reappraisal. Front. Psychol. 2. https://doi.org/10.3389/fpsyg.2011.00216.

Plichta, M.M., Schwarz, A.J., Grimm, O., Morgen, K., Mier, D., Haddad, L., Gerdes, A.B.M., Sauer, C., Tost, H., Esslinger, C., Colman, P., Wilson, F., Kirsch, P., Meyer-Lindenberg, A., 2012. Test–retest reliability of evoked BOLD signals from a cognitive–emotive fMRI test battery. NeuroImage 60, 1746–1758. https://doi.org/10.1016/j.neuroimage.2012.01.129.

Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., Munafò, M.R., Nichols, T.E., Poline, J.-B., Vul, E., Yarkoni, T., 2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat. Rev. Neurosci. 18, 115–126. https://doi.org/10.1038/nrn.2016.167.

Quiton, R.L., Keaser, M.L., Zhuo, J., Gullapalli, R.P., Greenspan, J.D., 2014. Intersession reliability of fMRI activation for heat pain and motor tasks. NeuroImage: Clinical 5, 309–321. https://doi.org/10.1016/j.nicl.2014.07.005.

Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R.J.A., Kahn, R.S., Ramsey, N.F., 2007. Test–retest reliability of fMRI activation during prosaccades and antisaccades. NeuroImage 36, 532–542. https://doi.org/10.1016/j.neuroimage.2007.03.061.

Reinhardt, T., Schmahl, C., Wüst, S., Bohus, M., 2012. Salivary cortisol, heart rate, electrodermal activity and subjective stress responses to the Mannheim Multicomponent Stress Test (MMST). Psychiatry Res. 198, 106–111. https://doi.org/10.1016/j.psychres.2011.12.009.

Scheier, M.F., Carver, C.S., 1985. Optimism, coping, and health: assessment and implications of generalized outcome expectancies. Health Psychol. 4.

Schwarzer, R., Jerusalem, M. (Eds.), 1999. Skalen zur Erfassung von Lehrer- und Schülermerkmalen. Dokumentation der psychometrischen Verfahren im Rahmen der Wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen. Freie Universität, Berlin.

Sebastian, A., Pohl, M.F., Klöppel, S., Feige, B., Lange, T., Stahl, C., Voss, A., Klauer, K.C., Lieb, K., Tüscher, O., 2013. Disentangling common and specific neural subprocesses of response inhibition. NeuroImage 64, 601–615. https://doi.org/10.1016/j.neuroimage.2012.09.020.

Sheehan, D.V., Lecrubier, Y., Sheehan, K.H., Amorim, P., Janavs, J., Weiller, E., Herqueta, T., Baker, R., Dunbar, G.C., 1998. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J. Clin. Psychiatry 59, 22–33.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass Correlations : uses in assessing rater reliability. Psychol. Bull. 86, 420–428.

Smith, B.W., Dalen, J., Wiggins, K., Tooley, E., Christopher, P., Bernard, J., 2008. The brief resilience scale: assessing the ability to bounce back. Int. J. Behav. Med. 15, 194–200. https://doi.org/10.1080/10705500802222972.

Taylor, S., Zvolensky, M.J., Cox, B.J., Deacon, B., Heimberg, R.G., Ledley, D.R., Abramowitz, J.S., Holaway, R.M., Sandin, B., Stewart, S.H., Coles, M., Eng, W., Daly, E.S., Arrindell, W.A., Bouvard, M., Cardenas, S.J., 2007. Robust dimensions of anxiety sensitivity: development and initial validation of the Anxiety Sensitivity Index-3. Psychol. Assess. 19, 176–188. https://doi.org/10.1037/1040-3590.19.2.176.

Turner, B.O., Paul, E.J., Miller, M.B., Barbey, A.K., 2018. Small sample sizes reduce the replicability of task-based fMRI studies. Commun. Biol. 1. https://doi.org/10.1038/s42003-018-0073-z.

Wu, C.C., Samanez-Larkin, G.R., Katovich, K., Knutson, B., 2014. Affective traits link to reliable neural markers of incentive anticipation. NeuroImage 84, 279–289. https://doi.org/10.1016/j.neuroimage.2013.08.055.