Test–retest reliability of neural alcohol cue-reactivity: Is there light at the end of the magnetic resonance imaging tube?

Patrick Bach1,2 | Iris Reinhard3 | Anne Koopmann1,2 | Jan M. Bumb1,2 | Wolfgang H. Sommer1,4 | Sabine Vollstädt-Klein1 | Falk Kiefer1,2

1Department of Addictive Behavior and Addiction Medicine, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany
2Feuerlein Center on Translational Addiction Medicine (FCTS), University of Heidelberg, Heidelberg, Germany
3Department of Biostatistics, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany
4Institute of Psychopharmacology, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

Correspondence
Patrick Bach, Department of Addictive Behavior and Addiction Medicine, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, J5/68159 Mannheim, Germany.
Email: patrick.bach@zi-mannheim.de

Funding Information
Deutsche Forschungsgemeinschaft, Grant/Award Number: 402170461-TRR 265

Abstract
Over the last decades, the assessment of alcohol cue-reactivity gained popularity in addiction research, and efforts were undertaken to establish neural biomarkers. This attempt however depends on the reliability of cue-induced brain activation. Thus, we assessed test–retest reliability of alcohol cue-reactivity and its implications for imaging studies in addiction. We investigated test–retest reliability of alcohol cue-induced brain activation in 144 alcohol-dependent patients over 2 weeks. We computed established reliability estimates, such as intraclass correlation (ICC), Dice and Jaccard coefficients, for the three contrast conditions of interest: ‘alcohol’, ‘neutral’ and the ‘alcohol versus neutral’ difference contrast. We also investigated how test–retest reliability of the different contrasts affected the capacity to establishing associations with clinical data and determining effect size estimates. Whereas brain activation, indexed by the constituting contrast conditions ‘alcohol’ and ‘neutral’ separately, displayed overall moderate (ICC > 0.4) to good (ICC > 0.75) test–retest reliability in areas of the mesocorticolimbic system, the difference contrast ‘alcohol versus neutral’ showed poor overall reliability (ICC < 0.40), which was related to the intercorrelation between the constituting conditions. Data simulations and analyses of craving data confirmed that the low reliability of the difference contrast substantially limited the capacity to establish associations with clinical data and precisely estimate effect sizes. Future research on alcohol cue-reactivity should be cautioned by the low reliability of the common ‘alcohol versus neutral’ difference contrast. We propose that this limitation can be overcome by using the constituent task conditions as an individual difference measure, when intending to longitudinally monitor brain responses.

Keywords
alcohol cue-reactivity, Dice, fMRI, intraclass correlation, Jaccard, reliability
1 | INTRODUCTION

The use of functional magnetic resonance imaging (fMRI)-based tasks gained popularity in addiction research over the last decade. Recently, the advent of research domain criteria (RDoC) has led to a surge of interest in individual biomarkers for mental disorders that include neural biomarkers. \(^1\,^2\) Particularly, the investigation of neural responses to addiction-related imagery was used to determine the neurobiological underpinnings of the individual reactivity to drug cues and associated behaviours, such as craving and relapse. \(^3\,^4\) Although the so-called ‘cue-reactivity’ paradigms were established for almost every stimulus category and over 100 publications reported fMRI data from such paradigms, to our knowledge, only a single study in \(n = 9\) alcohol-dependent patients specifically assessed the reliability and temporal stability of fMRI alcohol cue-reactivity tasks over a period of 14 days in the ventral striatum (VS) and dorsal striatum (DS). \(^7\) This seems surprising, because reliability is a prerequisite for the aim of many imaging studies, that is, associating brain activation with behavioural variables (e.g., craving), predicting future behaviour and developing treatment responsive biomarkers and neuroscience-based treatments for alcohol addiction. \(^8\,^9\)

A recent meta-analysis pointed out that the overall reliability of task fMRI across different tasks was poor. \(^10\) This might be due to the fact that in many fMRI studies, difference scores or difference contrasts between two constituting task conditions are computed. For example, for alcohol cue-reactivity tasks, it is a common procedure to subtract the brain activation during alcohol picture blocks from activation during neutral picture blocks. However, in the case of a high correlation between the constituting conditions, the resulting reliability of the resulting difference score is limited, because much of the shared ‘true’ variance is removed. \(^11\,^12\) Such a measure largely consists of unsystematic error variance, which would not show robust and replicable associations with any external variable and whose effects would not replicate. Even though alcohol cue-reactivity paradigms were implemented in many studies, no study to date assessed whole-brain reliability of this paradigm and its implications.

Hence, we set out to assess the reliability of the difference contrast alcohol versus neutral that is commonly computed in alcohol cue-reactivity magnetic resonance imaging (MRI) studies of an established alcohol cue-reactivity task and its implications for establishing associations with clinical data and estimating treatment effects. We assumed that the alcohol and neutral contrast conditions separately show higher test–retest reliability compared with the difference contrast alcohol versus neutral. Specifically, we hypothesized that the overlap of significantly activated voxels between first and second fMRI assessments, indexed by the Jaccard and Dice coefficients, would be higher for the neutral and alcohol contrast conditions, compared with the difference contrast. Further, we hypothesized that the magnitude of voxel-wise activation shows a higher resemblance between fMRI sessions, expressed by the intraclass correlation (ICC) coefficient, for the neutral and alcohol contrast conditions, compared with the difference contrast. Moreover, we expected that a higher proportion of patients could be re-identified by their neural activation patterns from first to second fMRI, expressed by a higher within-subject similarity, for the neutral and alcohol contrast conditions, compared with the difference contrast.

2 | METHODS AND MATERIALS

2.1 | Study sample and patient subgroups

Datasets of \(N = 144\) alcohol-dependent patients were included in the current analyses. Patients were recruited as part of three studies (NCT01503931, NCT00926900 and DRKS00003357). All procedures were carried out in accordance to the Declaration of Helsinki, and the local ethics committee approved the study procedures, and all participants provided informed written consent. All patients were abstinent at the time of MRI assessment and remained in in-patient or day-care treatment at the Central Institute of Mental Health (Mannheim) during the time between the first and second MRI sessions. Abstinence from substance use was controlled by daily breath alcohol testing and random drug urine screening. Data of five patient subgroups (see Figure 1) were included in the current analyses. All patients received treatment as usual (i.e., multidisciplinary 3-week intensified withdrawal treatment [IWT]).

Subgroup 1: The first subgroup of \(n = 21\) patients (titled IWT subgroup) received treatment as usual (i.e., a treatment programme that runs about 21 days with a daily multi-professional medically supervised therapy schedule, here termed intensified withdrawal treatment). \(^13\)

Subgroup 2: The second subgroup of \(n = 42\) patients (IWT + CET I group) also received standard in-patient treatment and individualized cue-exposure treatment (CET) sessions (five to nine sessions per patients of 60 to 90 min each) between baseline and second fMRI sessions. Patients received multiple sessions of CET; that is, they were exposed to their favourite drink (i.e., viewing handling and smelling, but no consumption), while being supervised by a trained psychologist until the cue-induced craving returned to zero. In contrast to the IWT + CET II group, a different version of the alcohol cue-reactivity task was included in this study that did not include a rating phase during the task (see section on fMRI task design).

Subgroup 3: The third subgroup (IWT + NTX) consists of \(n = 20\) patients that received IWT plus adjuvant oral naltrexone (NTX, 50 mg/day). NTX treatment was initiated after the baseline fMRI and continued until the second fMRI that was conducted at about 14 days into treatment. Detailed results are reported elsewhere. \(^3\)

Subgroup 4: The fourth subgroup (IWT + CET II group) was \(n = 29\) patients that received about nine CET sessions in addition to IWT. The details of the study are described in detail in a previous publication. \(^14\)
Subgroup 5: The fifth subgroup of $n=32$ patients (IWT + CET + DCS group) underwent standard in-patient treatment and nine standardized CET sessions and additionally received a 50-mg dose of D-cycloserine (DCS), a partial agonist at the glycine binding site of the NMDA receptor, 1 h prior to each CET session. Subgroup 5 received nine doses of 50-mg dose of D-cycloserine (DCS) concurrently with nine CET sessions over 2 weeks in addition to IWT. Two versions of an established alcohol picture cue-reactivity task were used (ALCUEPV = Alcohol Cue-Reactivity Picture Viewing Task and ALCUE = Alcohol Cue-Reactivity Task). Both versions of the alcohol cue-reactivity task use a block design that includes the presentation of series of either alcohol or neutral pictures in subsequent blocks that were presented in a pseudo-randomized order. In contrast to the ALCUE task, the ALCUEPV task did not include a rating phase in-between successive picture blocks.

Detailed exclusion and inclusion criteria and subgroup characteristics are reported in the supplementary Methods section (see Supplementary Methods).

2.2 | Assessment

All patients underwent two assessment sessions, both including fMRI of alcohol cue-reactivity. The first assessment was scheduled prior to initiation of any intervention in addition to IWT.

2.3 | fMRI alcohol cue-reactivity tasks, fmRI acquisition and pre-processing

Two versions of an established alcohol picture cue-reactivity task were used. The tasks were validated in previous studies. Both versions of the alcohol cue-reactivity task use a block design that includes the presentation of series of either alcohol or neutral pictures in subsequent blocks. Per block, series of five alcohol-related or neutral pictures were presented to participants via MRI compatible googles (MRI Audio/Video Systems, Resonance Technology Inc., Los Angeles, CA, USA) in a pseudo-randomized order. In the first version of the task (referred to as ALCUE = Alcohol Cue-Reactivity Task), a total of 12 blocks of alcohol pictures (five pictures per block) and nine blocks of neutral pictures (five pictures per block) were displayed. Each picture was presented for 4 s. In the first version of the task, participants were asked to rate their current subjective craving after each block (e.g., 21 times) on a visual analogue scale from 0 (no craving at all) to 100 (very intense craving). The task took approximately 12 min to complete in its entirety, depending on the duration of the rating phases (i.e., time till participants entered their response).

In the second version of the task (referred to as ALCUEPV = Alcohol Cue-Reactivity Passive Viewing Task), there was no rating phase, and the task consisted of 12 blocks of both alcohol stimuli and 12 blocks of the same neutral stimuli that were presented for 4 s each. The task was designed, in order to keep the time on task similar across participants (i.e., avoid rating phases that could be individually longer or shorter, hence affecting the delay between blocks and the overall time on task). This task took 12:15 min. Data
presentation and recording was monitored using the Presentation software (Version 16.0, Neurobehavioral Systems Inc., Albany, CA, USA).

Detailed information on the fMRI acquisition and pre-processing procedures are presented in the Supplements (see Supplementary Methods).

### 2.4 | Reliability measures

All reliability analyses were conducted using the SPM Reliability Toolbox (https://github.com/CPernet/spmrt/) by Cyril Pernet and colleagues and the fmreli toolbox for SPM12 (https://github.com/nkroemer/reliability). Individual contrast images of the different task conditions served as input for the reliability analyses.

#### 2.4.1 | Intraclass correlation

Reliability of fMRI data was estimated for each voxel of the brain activation maps by computing the ICC coefficients between time points. It was argued that the ICC(3,1) type is most appropriate for assessing longitudinal fMRI data. Hence, the ICC(3,1) type was computed (for details, see Supplementary Methods) for the difference contrast alcohol versus neutral and the neutral and alcohol contrasts separately. Thresholded ICC brain maps were created, in order to identify brain areas that show moderate to good (ICC > 0.4) and good to excellent (ICC > 0.75) reliability. To facilitate the assessment of local differences in reliability, we computed the mean ICC for anatomical regions specified in the Automatic Anatomic Labeling (aal) atlas. For details, see Supplementary Methods.

#### 2.4.2 | Jaccard and Dice coefficients

The Jaccard and Dice coefficients were computed for every participant to investigate overlap of significant voxels surpassing a predefined statistical threshold between the first and second fMRI (see Supplementary Methods). Repeated measures analyses of variance models with the factors contrast category (alcohol, neutral and alcohol vs. neutral) and groups (five patient subgroups) were used to test main effects and interactions on the magnitude of the Jaccard and Dice coefficients.

#### 2.4.3 | Similarity

We calculated the similarity of the fMRI activation maps using the fmreli toolbox. This procedure captures the resemblance of two activation patterns based on the alignment of high versus low brain activation values across the brain. It was suggested that subjects can be re-identified by their neural activation patterns, if the within-subject similarity exceeds all between-subject association coefficients of the same participant.

#### 2.4.4 | Correlation between contrast maps

In order to assess the correlation between the different task conditions, voxel-wise Pearson correlation coefficients were computed between the alcohol and neutral contrast conditions, as well as between the alcohol versus neutral difference contrast and the former two conditions using the fmreli toolbox.

#### 2.4.5 | Analyses of group-level fMRI activation

On a group level, imaging data were analysed using full factorial models with the factor time (first and second scan) for the five separate subgroups, in order to assess the congruence and robustness of task main effects on the group level brain activation over time (contrast: alcohol vs. neutral). We conducted additional analyses of first-level contrast images (contrast: alcohol vs. neutral) by applying a repeated measures full factorial design using the SPM12 software toolbox with factors study subgroup (n =5) as between subject factor and time (T1, T2) as within-subject factor. In order to satisfy a family-wise error rate correction of $p_{FWE} < 0.05$, we determined a combined voxel-wise ($p < 0.001$) and cluster-extent threshold ($k ≥ 110$) by running 10,000 permutations by Monte Carlo simulations (the estimated smoothness was $x/y/z = 10.22/10.75/9.03\ mm$) using the Neuroelf analysis package (www.neuroelf.net).

#### 2.4.6 | Analyses of associations with clinical data and simulation analyses

We investigated associations between subjective craving, using the Obsessive Compulsive Drinking Scale (OCDS), and brain activation in the two regions of interest (ROIs) (putamen and caudate), which showed moderate to good reliability in the reliability analyses, for the neutral, alcohol and difference (alcohol vs. neutral) contrasts (see Supplementary Methods). Additionally, we performed simulation analysis to test our hypothesis that associations between the difference contrast and clinical variables are limited by the magnitude of the intercorrelation between the constituting contrast conditions. Simulation was carried out for determining the estimated correlation between the difference score of the two constituting variables with external variables for varying correlation parameters. In doing this, samples of size $N = 144$ were generated (see Supplementary Methods). Due to fact that the reliability of fMRI brain activation attenuates the observed effect size, compared with the population parameter, we assessed how test-retest reliability (or lack thereof) would attenuate small, medium and large effect sizes resulting from clinical studies (see Supplementary Methods).
3 | RESULTS

3.1 | Sample characteristics

The patient subgroups showed similar values across demographical and psychometric variables with the exception of the Beck Depression Inventory (BDI) scores and time of abstinence (see Table 1). The mean abstinence of patients at baseline was 13.1 days (SD = 8.8) and 29.7 days (SD = 9.9) at the time of the second assessment (mean difference 16.7 days [SD = 4.3]). None of the patients relapsed between the baseline and second assessment. Comparison between fMRI sessions showed a significant effect of time for the magnitude of alcohol craving, measured using the Alcohol Urge Questionnaire (AUQ) ($F = 15.20, p < 0.001$) and the OCDS ($F = 60.61, p < 0.001$) and a significant interaction between study subgroup and time for the OCDS ($F = 8.69, p = 0.049$), whereas there was no significant interaction between study subgroup and time for the AUQ ($F = 2.91, p = 0.146$) and no significant main effect of subgroup on the magnitude of AUQ ($F = 1.317, p = 0.273$) or OCDS ($F = 0.99, p = 0.397$) scores. The significant interaction effect was driven by a more pronounced craving reduction, measured using the OCDS, in the subgroup receiving IWT and NTX, compared with the other study subgroups (post hoc tests: $p < 0.044$, see Table 1).

3.2 | Robust main effects of the alcohol versus neutral contrast

Analyses of group-level brain activation demonstrated robust alcohol cue-induced brain response (contrast: alcohol vs. neutral) across the different subgroups and for both fMRI assessment sessions including the inferior, middle and superior occipital gyri, the cuneus, lingual and fusiform gyri as well as middle and superior parietal gyri, the putamen and caudate (see Table S1). There was no significant main effect of time or subgroup on the magnitude of alcohol cue-induced brain activation (contrast: alcohol vs. neutral).

3.3 | Moderate to good mean test–retest reliability of the alcohol and neutral contrast conditions

ICC values for the alcohol contrast condition showed moderate mean global ICC values, when computing the mean ICC across the whole brain ($\text{ICC}_{\text{Mean}} > 0.40$, see Table 2) and good to excellent local reliability ($\text{ICC} > 0.75$), mainly in the occipital gyri and to a lesser extent in the inferior and superior frontal gyri (see Figure 2) and moderate to good reliability ($\text{ICC} > 0.4$) in large clusters of brain areas including mesolimbic brain regions (insula, putamen and caudate), as well as temporal and parietal gyri (see Figure S1). The patterns of brain areas depicting moderate and good reliability for the neutral contrast condition resembled these findings (see Table 2).

Similarity analyses showed higher within-subject similarity compared with between-subject similarity over time for the alcohol and neutral contrast conditions ($\text{ICC}_{\text{Alcohol}} \geq 11.02, p < 0.001, \text{ICC}_{\text{Neutral}} \geq 9.85, p < 0.001$) with overall high within-subject similarity values ($\text{ICC}_{\text{Alcohol}} = 0.71, \text{ICC}_{\text{Neutral}} = 0.69$), which is visible in a prominent diagonal in the similarity matrices of the alcohol and neutral contrast maps (see Figure 3). This translated into the observation that >77% of the individual patients could be re-identified, based on their neural activation pattern during alcohol and neutral contrast conditions.

The Dice and Jaccard coefficients showed values ranging from 0.44 to 0.64 for the alcohol and neutral contrasts, indicating that about half of the significant activation clusters could be replicated in the second fMRI session (see Table S3A,B). Across both statistical thresholds, both overlap indices were significantly higher for the alcohol and neutral contrast conditions than for the alcohol versus neutral difference contrast ($F_{2,278} = 10.717, p < 0.001$, for detailed results, see Supplementary Results).

3.4 | Poor mean test–retest reliability of the alcohol versus neutral difference contrast

In contrast to the robust main effects and contrary to the good reliability of the alcohol and neutral contrasts separately, the difference contrast showed poor mean whole-brain reliability (ICC$_{\text{Mean}} < 0.4$) (see Table 2A). Subsequent voxel-wise analyses using thresholded ICC maps showed that several brain areas surpassed the threshold of ICC > 0.75 (indicating good reliability) in the IWT and IWT + NTX study groups. These areas included the bilateral insulae, part of the rostral right putamen, parts of the inferior medial occipital gyrus, parts of the bilateral middle and inferior frontal gyri (specifically the Heschl gyri), the left cuneus and parts of the right lateral frontal gyrus and bilateral superior frontal gyri (see Figure 4). Further analyses indicated that several additional brain areas showed moderate to good reliability, such as the occipital gyrus, parts of the putamen and caudate and parts of the inferior and superior frontal gyri (see Figure S1A–E).

Similarity analyses showed that overall similarity values for the difference contrast were low ($\text{ICC}_{\text{Alcohol–Neutral}} = 0.19$) and within-subject similarity did not exceed between-subject similarity ($\text{ICC}_{\text{Alcohol–Neutral}} \leq 1.83, p > 0.05$) (see Figure 3). This reflected in a low proportion of patients of 22% that could be re-identified based on their neural activation signature captured by the difference contrast (alcohol vs. neutral).

The Dice and Jaccard coefficients indicated minimal overlap of significant activation clusters between scanning time points of about 0.07 to 0.14, indicating that only a small fraction of super-threshold clusters could be replicated. The overlap indices were significantly lower for the alcohol versus neutral difference contrast (see Table S3A–B), compared with the constituting task conditions across both statistical thresholds ($p < 0.001, p < 0.01$) that were used to define significant voxels (Jaccard: $F_{2,278} = 860.093, p < 0.001$ and $F_{2,278} = 648.951, p < 0.001$, respectively; Dice: $F_{2,278} = 948.150, p < 0.001$ and $F_{2,278} = 10.717, p < 0.001$, respectively) (for detailed results, see Supplementary Results).
TABLE 1  Demographic data, alcohol use and severity measures for patient groups with available imaging data for both time points (baseline and Week 2 scan)

| Subgroup | 1 | 2 | 3 | 4 | 5 |
|----------|---|---|---|---|---|
| IWT (ALCUEPV) (n = 21) | IWT + CET I (ALCUEPV) (n = 42) | IWT + CET II (ALCUEPV) (n = 20) | IWT + CET + DCS (ALCUE) (n = 29) | IWT + CET + NTX (ALCUEPV) (n = 20) |
| Demographical variables | | | | | |
| Age (years) | 47.6 (10.5) | 46.6 (10.0) | 48.3 (8.4) | 47.5 (10.3) | 44.0 (10.1) |
| Education (no post-secondary educ./apprenticeship only/attended college/higher education) | 2/9/5/5 | 8/22/7/5 | 2/13/1/3 | 4/13/5/7 | 6/22/2/2 |
| Substance use patterns | | | | | |
| Ethanol (g/day; mean of last 90 days) | 178.8 (104.5) | 162.5 (129.0) | 182.3 (135.6) | 108.0 (96.2) | 126.7 (98.7) |
| Drinks per drinking day (mean of last 90 days) | 19.1 (9.4) | 18.4 (13.6) | 17.3 (10.3) | 13.5 (7.8) | 14.4 (8.4) |
| Abstinent days (% in last 90 days) | 22.0 (19.0) | 22.0 (28.1) | 16.1 (22.3) | 34.2 (33.6) | 26.8 (29.1) |
| Heavy-drinking days (% in last 90 days) | 74.6 (20.0) | 71.2 (30.9) | 76.2 (28.0) | 58.8 (34.6) | 63.7 (32.6) |
| Abstinence in days before baseline fMRI | 9.9 (6.0) | 12.5 (11.0) | 20.6 (7.3) | 12.1 (7.4) | 12.0 (6.8) |
| Abstinence in days before 2nd fMRI | 25.2 (7.3) | 28.2 (7.9) | 36.2 (7.8) | 28.3 (8.6) | 30.3 (8.7) |
| Days between 1st and 2nd fMRI sessions | 16.1 (3.6) | 16.9 (3.4) | 15.6 (3.6) | 16.2 (4.7) | 18.2 (4.4) |
| Smoker (yes/no) | 10:11 | 27:15 | 13:7 | 21:8 | 24:8 |
| Cigarettes per day (smokers only) | 1.7 (1.2) | 1.6 (1.0) | 1.1 (1.3) | 1.6 (1.1) | 1.3 (0.9) |
| Clinical scales | | | | | |
| OCDS (total score) | 18.0 (7.6) | 15.5 (7.0) | 13.7 (5.8) | 16.1 (6.0) | 15.9 (6.5) |
| FTND (total score) | 3.7 (3.8) | 3.7 (3.3) | 5.5 (2.7) | 5.0 (3.2) | 4.7 (2.9) |
| ADS (total score) | 15.7 (6.3) | 16.2 (6.7) | 12.9 (5.9) | 16.6 (5.7) | 14.8 (7.4) |
| STAI (trait sum score) | 44.2 (15.9) | 45.8 (12.0) | 37.9 (10.8) | 43.2 (9.6) | 46.3 (10.6) |
| STAI (state sum score) | 40.2 (12.8) | 43.5 (12.0) | 36.9 (8.3) | 43.4 (10.6) | 45.0 (10.1) |
| AUQ (total score) | 12.9 (5.3) | 12.6 (5.8) | - | 11.2 (4.9) | 134 (7.4) |
| BDI (total score) | 14.5 (11.7) | 15.9 (10.1) | 9.7 (8.0) | 10.3 (7.4) | 11.6 (8.9) |
| 2nd fMRI—Clinical scales | | | | | |
| OCDS (total score) | 12.2 (6.8) | 9.9 (5.9) | 4.0 (4.1) | 11.6 (5.7) | 10.8 (5.3) |
| AUQ (total score) | 10.3 (3.4) | 9.9 (3.1) | - | 9.3 (2.4) | 10.9 (3.4) |

Abbreviations: ADS, Alcohol Dependence Scale; ALCUE, Version 1 of the alcohol cue-reactivity task; ALCUEPV, Version 2 of the alcohol cue-reactivity task without a rating phase (see Supplementary Methods for details); BDI, Beck Depression Inventory; fMRI, functional magnetic resonance imaging; FTND, Fagerstrom Test for Nicotine Dependence; IWT, subgroup receiving treatment as usual (i.e., multidisciplinary 3-week intensified withdrawal treatment [IWT]); IWT + CET I, subgroup receiving nine cue-exposure treatment (CET) sessions in addition to IWT; IWT + CET II, subgroup receiving nine cue-exposure treatment (CET) sessions in addition to IWT; IWT + CET + DCS, subgroup receiving nine doses of 50-mg dose of d-cycloserine (DCS) concurrently with nine CET sessions in addition to IWT; IWT + CET + NTX, subgroup receiving daily naltrexone (NTX) in addition to IWT; OCDS, Obsessive Compulsive Drinking Scale; STAI, State-Trait-Anxiety Inventory; SD, standard deviation.

Significant post hoc test *p < 0.05.

*Significant main effect of group *p < 0.05.
### TABLE 2  Mean ICC values across groups and contrast conditions

| Group | IWT (ALCUEPV) \(n=21\) Mean (SD) \[range\] | IWT + CET II (ALCUE) \(n=29\) Mean (SD) \[range\] | IWT + CET I (ALCUEPV) \(n=42\) Mean (SD) \[range\] | IWT + CET + DCS (ALCUE) \(n=32\) Mean (SD) \[range\] | IWT + NTX (ALCUEPV) \(n=20\) Mean (SD) \[range\] |
|-------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
| A. Difference contrast: Alcohol–neutral | | | | | |
| Group \(n=21\) | Mean (SD) \[range\] | Mean (SD) \[range\] | Mean (SD) \[range\] | Mean (SD) \[range\] |
| IWT | 0.23 (0.23) \[0.20–0.70\] | 0.18 (0.18) \[0.23–0.63\] | 0.06 (0.06) \[0.57–0.69\] | 0.00 (0.00) \[0.73–0.67\] | 0.18 (0.19) \[0.79–0.84\] |
| IWT + CET II \(n=29\) | 0.18 (0.18) | 0.18 (0.18) | 0.06 (0.06) | 0.00 (0.00) | 0.18 (0.19) |
| IWT + CET I \(n=42\) | 0.06 (0.06) | 0.06 (0.06) | 0.00 (0.00) | 0.00 (0.00) | 0.00 (0.00) |
| IWT + CET + DCS \(n=32\) | 0.00 (0.00) | 0.00 (0.00) | 0.00 (0.00) | 0.00 (0.00) | 0.00 (0.00) |

| B. Contrast: Alcohol | | | | | |
| Group \(n=21\) | Mean (SD) \[range\] | Mean (SD) \[range\] | Mean (SD) \[range\] | Mean (SD) \[range\] |
| IWT | 0.40 (0.25) \[0.25–0.79\] | 0.54 (0.22) \[0.53–0.97\] | 0.30 (0.23) \[0.47–0.92\] | 0.43 (0.25) \[0.78–0.96\] | 0.38 (0.25) \[0.63–0.96\] |
| IWT + CET II \(n=29\) | 0.54 (0.22) | 0.54 (0.22) | 0.30 (0.23) | 0.30 (0.23) | 0.30 (0.23) |
| IWT + CET I \(n=42\) | 0.30 (0.23) | 0.30 (0.23) | 0.30 (0.23) | 0.30 (0.23) | 0.30 (0.23) |
| IWT + CET + DCS \(n=32\) | 0.43 (0.25) | 0.43 (0.25) | 0.43 (0.25) | 0.43 (0.25) | 0.43 (0.25) |

| C. Contrast: Neutral | | | | | |
| Group \(n=21\) | Mean (SD) \[range\] | Mean (SD) \[range\] | Mean (SD) \[range\] | Mean (SD) \[range\] |
| IWT | 0.25 (0.29) \[0.25–0.72\] | 0.46 (0.22) \[0.49–0.96\] | 0.27 (0.24) \[0.58–0.92\] | 0.21 (0.31) \[0.73–0.96\] | 0.39 (0.24) \[0.66–0.96\] |
| IWT + CET II \(n=29\) | 0.46 (0.22) | 0.46 (0.22) | 0.46 (0.22) | 0.46 (0.22) | 0.46 (0.22) |
| IWT + CET I \(n=42\) | 0.27 (0.24) | 0.27 (0.24) | 0.27 (0.24) | 0.27 (0.24) | 0.27 (0.24) |
| IWT + CET + DCS \(n=32\) | 0.21 (0.31) | 0.21 (0.31) | 0.21 (0.31) | 0.21 (0.31) | 0.21 (0.31) |

**Abbreviations:** ALCUE, Version 1 of the alcohol cue-reactivity task; ALCUEPV, Version 2 of the alcohol cue-reactivity task without a rating phase (see Supplementary Methods for details); IWT, subgroup receiving treatment as usual (i.e., multidisciplinary 3-week intensified withdrawal treatment [IWT]); IWT + CET II, subgroup receiving nine cue-exposure treatment (CET) sessions in addition to IWT; IWT + CET I, subgroup receiving nine cue-exposure treatment (CET) sessions in addition to IWT; IWT + CET + DCS, subgroup receiving nine doses of 50-mg dose of D-cycloserine (DCS) concurrently with nine CET sessions in addition to IWT; IWT + NTX, subgroup receiving daily naltrexone (NTX) in addition to IWT.

*p < 0.05. **p < 0.01. ***p < 0.005.*
The atlas-based summary of mean ICC values for the difference contrast alcohol versus neutral again showed that only a few brain regions surpassed a mean ICC value of 0.4 for moderate reliability. Specifically, the left superior occipital gyrus, the left and right Heschl gyri and the left cuneus showed values of 0.4 or higher in the IWT group that received treatment as usual (see Table S2). This pattern was not consistent across groups. Atlas-based summary measures for the Patient Subgroups 2, 3 and 4 did not exceed values of 0.4, except for the left superior occipital gyrus in the IWT + CET I subgroup (Group 2). The IWT + NTX subgroup showed mean ICC values > 0.4 for the bilateral calcarine and the left pallidum. With regard to the left and right DS and VS masks that were adapted from Schacht et al.7 to facilitate comparability, results show poor mean reliability (ICC < 0.4) across all groups in all of the four masks (see Table S2).

3.5 | Assessment of factors underlying the reliability differences across contrast conditions

Results show a high correlation between the alcohol and neutral task condition contrast maps ($r = 0.48$, SD = 0.34, $R^2 = 0.23$). This indicates that both conditions share about 23% of their variance. A part of this variance is removed by subtracting both conditions, illustrated by the lower correlation coefficients between the difference contrast and the constituting condition contrast maps (alcohol, $r = 0.24$, SD = 0.25, $R^2 = 0.06$; neutral, $r = -0.25$, SD = 0.24, $R^2 = 0.06$).

3.6 | Impact of low test–retest reliability of the difference contrast

Analyses indicated a significant correlation between OCDS scores (total score) and brain activation in the left caudate, indexed by the alcohol contrast ($r = 0.231$, $p = 0.005$, $p_{FDR} = 0.01$, $n = 144$). The significant correlation however could not be replicated using the difference contrast ($r = 0.077$, $p = 0.356$, $p_{FDR} = 0.356$, $n = 144$), even though the sample size was sufficient to detect even small to medium effects ($r \geq 0.20$) with a power of 78%. Further, the data simulations indicated that the mean correlations between the difference score ($V_1 - V_2$) and the external variables, which were modelled according to the actual cue-reactivity and OCDS data, were always substantially lower, by about one third, compared with the correlation between the constituting conditions ($V_1$ and $V_2$) and the external variables (see Table S4 and Figure S3). Additional data simulations showed that any effect size measure, which was determined using tools with poor to moderate reliability (i.e., ICC scores < 0.6), substantially overestimated the population effect sizes (see Figure S4).

4 | DISCUSSION

Here, we present the first whole-brain investigation of the reliability of an established alcohol cue-reactivity task and demonstrate how the reliability of the common difference contrast (alcohol vs. neutral)
FIGURE 3  Similarity maps (upper row) and empirical cumulative distribution functions (lower row—red lines: between-subject similarity, blue lines: within-subject similarity) for longitudinal comparisons (first and second fMRI sessions) for the three contrasts: (A) alcohol, (B) alcohol–neutral and (C) neutral. The diagonal of each colour matrix represents the within-subject similarity values. Re-identification of a subject is considered possible, if the within-subject similarity value (diagonal) exceeds all between-subject association coefficients of the same participant (i.e., similarity values in the respective row of the matrix). Higher within-subject similarity is also illustrated by a right shift of the cumulative density functions for the within-subject similarity values (blue lines) relative to the between-subject similarity (red lines) for the (A) alcohol and (C) neutral contrast maps, whereas the cumulative density functions overlapped for the (B) alcohol–neutral contrast.

FIGURE 4  Depiction of brain areas that show good to excellent reliability for the difference contrast alcohol–neutral (intraclass correlation [ICC] > 0.75) for two study groups (all other groups did not show ICCs > 0.75)

Brain regions with ICC values > 0.75 ("good") for the difference contrast „alcohol vs. neutral“

(A) Subgroup 1 - IWT, n = 21

(B) Subgroup 3 - NTX, n = 20
determines the capacity to establish associations with clinical data and estimate effect sizes from clinical trials.

4.1 | Moderate to good reliability of the constituting task contrasts

Overall, reliability measures indicated good to excellent reliability over roughly 14 days between the first and second fMRI sessions across the five different study subgroups for the alcohol and neutral contrast conditions, supporting the robustness of our findings. Computation of the voxel-wise ICC showed moderate to excellent reliability in several brain areas of the mesocorticolimbic system. Additionally, roughly half of the significant activation clusters overlapped between successive fMRI scans for the alcohol and neutral contrasts, and about 70% of the patients could be re-identified based on their neural signature captured by either the alcohol or neutral contrast. In addition, the Jaccard and Dice coefficients for the alcohol and neutral contrast conditions indicated that—depending on the threshold for defining active voxels—about 50% to 60% of the activation clusters could be replicated from the first to second fMRI. This supports the general potential of fMRI-based cue-reactivity measures to provide reliable indicators of individual brain activation in areas that are part of the addiction network.

4.2 | Poor reliability of the difference contrasts and its impact on imaging studies

In sharp contrast, the difference contrast (alcohol–neutral) showed poor overall reliability, indicated by low mean ICC values and low Jaccard and Dice coefficients. The lower reliability of the difference contrast results from the substantial intercorrelation of the constituting contrast conditions (alcohol, neutral), which eliminates parts of the shared variance and summates the error variance. This phenomenon was already demonstrated in previous work by Infantolino and colleagues. Their data on the difference contrast between face- and shape-matching trials of the well-established faces paradigm showed very poor reliability of the difference contrast, which was attributed to a high correlation (0.97) of the constituting task conditions. A recent meta-analysis on 56 independent fMRI studies concluded that only 46% of the reported reliability scores fell within the range of at least moderate reliability. A subsequent moderator analysis indicated that neither task type, task design, task length, test–retest interval, ROI type (i.e., structural vs. functional), nor sample type (i.e., healthy vs. clinical) significantly moderated reliability scores. Although this seems unexpected, it might point towards a fundamental problem, that is, that most of the studies applied difference score measures. A measure that constitutes largely of unsystematic variance would show no systematic variation depending on the aforementioned factors. Elliott and colleagues concluded that difference scores will always have lower reliability than their constituent scores or conditions. Besides all aforementioned limitations, several brain areas that are part of the mesocorticolimbic system showed moderate to good and good to excellent reliability in the reference group with treatment as usual (IWT) for the difference contrast alcohol–neutral, and also in the group receiving treatment with naltrexone (IWT + NTX). This suggests that the fMRI signal, captured by the difference contrast, shows heterogeneous but locally satisfactory reliability to monitor treatment effects and treatment efficacy.

4.3 | Factors underlying the low reliability of the difference contrast

We showed that the low mean test–retest reliability of the difference contrast (alcohol vs. neutral) resulted from the substantial intercorrelation between the constituting task conditions (alcohol and neutral). The low reliability impaired the capacity to establishing associations between fMRI and clinical data (i.e., OCDS scores). Our data simulations fortified this assumption. They showed a systematic reduction in the magnitude of the correlation coefficients between the difference contrast and external variables by about 34%, compared with the correlation with the constituent task conditions (i.e., alcohol and neutral) separately. This illustrates that elimination of the shared variance dramatically impacts on the capacity to replicate correlations between the constituting task conditions and external variables when using the difference contrast. One could argue that the difference contrast is the relevant contrast and any association should be exclusively investigated using this contrast. However, the fact that this measure largely consists of error variance, depending on the extent of intercorrelation of the constituent task conditions, argues against this approach. Any analysis would be prone to producing spurious results (i.e., correlation with the error variance component).

4.4 | Overcoming low reliability of the alcohol versus neutral difference contrast

To overcome the problems associated with computing difference contrasts, we suggest to use one of the constituent task conditions as an individual difference measure, given that it has adequate reliability. In the case of a linear association and high correlation between the constituting conditions, one task condition could substitute the other quite well, without losing information on the individual differences between participants. Translating this to the alcohol cue-reactivity task, we argue that the alcohol condition can be used to monitor alcohol cue-induced brain responses over time and capture the potential impact of treatment interventions on this parameter, because one would assume that changes in brain responses to alcoholic stimuli and associated changes in alcohol cue-induced subjective alcohol cravings would more likely be captured by the alcohol condition contrast. The neutral condition on the other hand can serve as an index for the stability and reliability
of cue-induced fMRI signal in general, which should not change due to alcohol addiction treatment (e.g., CET or anti-craving medication).

Hence, we suggest a two-step approach. Firstly, the specificity of the fMRI signal for the cognitive processes under investigation should be supported by either relying on robust meta-analysis that could inform about the role of a certain brain region in the respective fMRI task or could be established by investigating within-subject effects between the constituting task conditions first, in order to identify regions that activate differently under each condition. These ROIs could be used to restrict the following analyses. In a next step, brain activation during the alcohol condition could be used as a measure that reliably captures individual differences over time, and the neutral condition could serve as a measure for the stability of the cue-induced fMRI signal.

We applied this approach to our dataset; that is, we specified the putamen and caudate as ROIs based on reliability scores and their roles for the cognitive process under investigation, that is, alcohol craving. Applying this approach, we could show a significant correlation between OCDS scores and caudate activation during the alcohol contrast, which could not be shown, when relying on the difference contrast. This is also supported by the results of our simulation analyses, illustrating the relevance of reliability and intercorrelation of the task conditions for establishing associations with clinical data.

4.5 | Limitations

It could be argued that the inclusion of patients without any treatment might be favourable with regard to yielding optimal reliability. However, we strongly advocate for testing reliability under the conditions in which the actual task is applied. When intending to use neural brain response as biomarker for monitoring treatment response, reliability of this putative biomarker should be tested under the very same conditions. Moreover, withholding standard patient care from patients would contradict principles of good clinical practice in research. Two versions of an alcohol cue-reactivity task were used, differing in that one of the versions included a rating phase between picture blocks and the second version did not. The version without a rating phase was designed to keep task duration constant across subjects and to minimize the risk of carry-over effects of the rating phase on the measured blood oxygenation level dependent (BOLD) response. However, at the group level, the comparison of task versions currently showed no significant difference in neural activation patterns between the two task versions. A possible explanation for this finding is that the rating phase was relatively short in relation to the duration of the stimulus blocks (<50%) and that possible effects were levelled out by averaging over the stimulus blocks. Furthermore, there was also no systematic advantage of either task version across study groups in terms of the reliability achieved. Based on the present data, both task versions appear to be similarly suited to investigate the neural response to alcohol stimuli.

5 | CONCLUSION

For the first time, we conducted longitudinal whole-brain analyses of test–retest reliability of an established alcohol cue-reactivity task and demonstrated that the low mean reliability of the difference contrast (alcohol vs. neutral) substantially limits the capacity to establish associations with clinical data and determine precise effect size estimates. Contrary to the low mean reliability of the difference contrast, the constituting task conditions showed good to excellent reliability. This disparity resulted from the substantial correlation between the constituting task conditions, which limited the shared ‘true’ variance of the difference contrast. This highlights the general conceptual problems associated with computing difference scores in alcohol cue-reactivity research. Still, the good test–retest reliability of the constituting task conditions supports the general potential of fMRI for providing reliable measures of brain activation. Our data simulations and association analyses also show that measures can be taken to overcome the problems that are associated with low reliability of the difference contrast. Future research on neural alcohol cue-reactivity should be cautioned by these findings and employ methods to overcome these limitations.

ACKNOWLEDGEMENTS

We thank Ms. U. Schmid for proofreading and editing the manuscript. This study was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project ID 402170461-TRR 265. The funding source had no role in the collection, analysis and interpretation of data, in the writing of the report or in the decision to submit the article for publication.

Open Access funding enabled and organized by Projekt DEAL.

WOA Institution: ZENTRALINSTITUT FUR SEELISCHE GESUNDHEIT

Blended DEAL: Projekt DEAL

CONFLICT OF INTEREST

None of the authors report any competing interests.

AUTHOR CONTRIBUTIONS

FK, PB and SVK were responsible for the study concept. PB, SVK, AK, JMB and WHS contributed to the acquisition of the data. PB, IR and SVK conducted the data analysis and interpretation of findings. PB, FK, AK, JMB and WHS drafted the manuscript. WHS, SVK, FK, AK, JMB and IR provided critical revision of the manuscript for important intellectual content. All authors critically reviewed the content and approved the final version for publication.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on qualified request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions, and individual data cannot be shared.

ORCID

Patrick Bach https://orcid.org/0000-0001-5962-019X
Anne Koopmann https://orcid.org/0000-0002-7261-9334
Falk Kiefer https://orcid.org/0000-0001-7213-0398
REFERENCES

1. Insel T, Cuthbert B, Garvey M, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am Psychiatric Assoc. 2010;4(7):748-751.

2. Voon V, Grodin E, Mandali A, et al. Addictions Neuroimaging Assessment (ANIA): towards an integrative framework for alcohol use disorder. Neurosci Biobehav Rev. 2020;113:492-506.

3. Bach P, Weil G, Pompili E, et al. Incubation of neural alcohol cue reactivity after withdrawal and its blockade by naltrexone. Addict Biol. 2020;25(1):e12717. https://doi.org/10.1111/adb.12717

4. Mann K, Vollstädt-Klein S, Reinhard I, et al. Predicting naltrexone response in alcohol-dependent patients: the contribution of functional magnetic resonance imaging. Alcohol Clin Exp Res. 2014;38(11):2754-2762.

5. Schacht JP, Randall PK, Latham PK, et al. Predictors of naltrexone response in a randomized trial: reward-related brain activation, OPRM1 genotype, and smoking status. Neuropsychopharmacology. 2017;42(13):2640-2653.

6. Grüsser SM, Wrase J, Klein S, et al. Cue-induced activation of the striatum and medial prefrontal cortex is associated with subsequent relapse in abstinent alcoholics. Psychopharmacology (Berl). 2004;175(3):296-302.

7. Schacht JP, Anton RF, Randall PK, Li X, Henderson S, Myrick H. Stability of fMRI striatal response to alcohol cues: a hierarchical linear modeling approach. Neuroimage. 2011;56(1):61-68.

8. Heilig M, Augier E, Pfarr S, Sommer WH. Developing neuroscience-based treatments for alcohol addiction: a matter of choice? Transl Psychiatry. 2019;9(1):255-266.

9. Heilig M, Sommer WH, Spanagel R. The need for treatment responsive translational biomarkers in alcoholism research. Curr Top Behav Neurosci. 2016;28:151-171.

10. Elliott ML, Knodt AR, Ireland D, et al. What is the test-retest reliability of common task-fMRI measures? New empirical evidence and a meta-analysis. bioRxiv. 2019:681700.

11. Peter JP, Churchill GA Jr, Brown TJ. Caution in the use of difference between-subjects comparisons. Neuroimage. 2018;173:146-152.

12. Infantolino ZP, Lukić KR, Sauder CL, Curtin JJ, Hajcak G. Robust is not necessarily reliable: from within-subjects fMRI contrasts to between-subjects comparisons. Neuroimage. 2018;173:146-152.

13. Loeber S, Kiefer F, Wagner F, Mann K, Croissant B. Treatment outcome after inpatient alcohol withdrawal: impact of motivational interventions: a comparative study. Nervenarzt. 2009;80(9):1085-1092.

14. Vollstädt-Klein S, Loeber S, Kirsch M, et al. Effects of cue-exposure treatment on neural cue reactivity in alcohol dependence: a randomized trial. Biol Psychiatry. 2011;69(11):1060-1066.

15. Kiefer F, Kirsch M, Bach P, et al. Effects of D-cycloserine on extinction of mesolimbic cue reactivity in alcoholism: a randomized placebo-controlled trial. Psychopharmacology (Berl). 2015;232(13):2353-2362.

16. Vollstädt-Klein S, Loeber S, Richter A, et al. Validating incentive salience with functional magnetic resonance imaging: association between mesolimbic cue reactivity and attentional bias in alcohol-dependent patients. Addict Biol. 2012;17(4):807-816.

17. Frohner JH, Teckentrup V, Smolka MN, Kroemer NB. Addressing the reliability fallacy in fMRI: similar group effects may arise from unreliable individual effects. Neuroimage. 2019;195:174-189.

18. Ombao H, Lindquist M, Thompson W, Aston J. Handbook of Neuroimaging Data Analysis. CRC Press; 2016.

19. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage. 2002;15(1):273-289.

20. Finn ES, Shen X, Scheinost D, et al. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat Neurosci. 2015;18(11):1664-1671.

21. Bennett CM, Wolford GL, Miller MB. The principled control of false positives in neuroimaging. Soc Cogn Affect Neurosci. 2009;4 (4):417-422. https://doi.org/10.1093/socan/asp053

22. Hunter JE, Schmidt FL. Correcting for sources of artificial variation across studies. Handbook Res Synth. 1994;236:323-336.

23. Hariri AR, Tessitore A, Mattay VS, Fera F, Weinberger DR. The amygdala response to emotional stimuli: a comparison of faces and scenes. Neuroimage. 2002;17(1):317-323.

24. Noori HR, Cosa Linan A, Spanagel R. Largely overlapping neuronal substrates of reactivity to drug, gambling, food and sexual cues: a comprehensive meta-analysis. Eur Neuropsychopharmacol. 2016;26(9):1419-1430.

25. Heinz A, Kiefer F, Smolka MN, et al. Addiction Research Consortium: losing and regaining control over drug intake (ReCoDe)—from trajectories to mechanisms and interventions. Addict Biol. 2020;25(2):e12866.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Bach P, Reinhard I, Koopmann A, et al. Test–retest reliability of neural alcohol cue-reactivity: Is there light at the end of the magnetic resonance imaging tube? Addiction Biology. 2021:e13069. https://doi.org/10.1111/adb.13069