Aus dem Zentralinstitut für Seelische Gesundheit

der Medizinischen Fakultät Mannheim

(Direktor: Prof. Dr. med. Andreas Meyer-Lindenberg)

Klinik für Abhängiges Verhalten und Suchtmedizin

(Direktor: Prof. Dr. med. Falk Kiefer)

Investigation of the reliability of fMRI-based assessments of neural cue-reactivity

Inaugural dissertation

zur Erlangung des Doctor scientiarum humanarum (Dr. sc. hum)

der Medizinischen Fakultät Mannheim

der Ruprecht-Karls-Universität

zu

Heidelberg

vorgelegt von

PD Dr. med. B.Sc. Patrick Bach

aus

Weinheim

2022

Dekan: Herr Prof. Dr. med. Sergij Goerdt

Referent: Herr Prof. Dr. med. Falk Kiefer

TABLE OF CONTENT

Page

LIST OF FIGURES

Main Figures

Supplementary Figures

LIST OF TABLES

Main Tables

Supplementary Tables

ABBREVIATIONS

AAC	anterior cingulate cortex
ADS	Alcohol Dependence Scale
AICA	Assessment of Internet & Computer Addiction
ALCUE	Alcohol cue-reactivity task
ALCUEPV	Alcohol cue-reactivity task with passive viewing
AUD	Alcohol Use Disorder
AUDIT	Alcohol Use Disorders Identification Test
AUQ	Alcohol Urge Questionnaire
BDI	Beck Depression Inventory
BLA	basolateral amygdala
BMI	Body Mass Index
BOLD	Blood Oxygenated Level Dependent
CeN	nucleus centralis of the amygdala
CET	cue-exposure treatment
CNR	contrast to noise ratio
CR	cue-reactivity
DCS	d-cycloserine
DS	dorsal striatum
DSM IV	Diagnostic and Statistical Manual of Mental Disorders, 4^{th} revision
fMRI	functional magnetic resonance imaging
FTND	Fagerström Test for Nicotine Dependence

FWE	family-wise error rate
ICC	intraclass correlation
ICD 10	International Classification of Diseases, 10th revision
IFG	inferior frontal gyrus
IDG	internet gaming disorders
IWT	intensified withdrawal treatment
HiTOP	Hierarchical Taxonomy of Psychopathology
mg	milligram
yg	microgram
ml	milliliter
MDD	Major Depressive Disorder
MRI	magnetic resonance imaging
NAc	Nucleus accumbens
NTX	naltrexone
ng	nanogram
OCDS	Obsessive and Compulsive Drinking Scale
OFC	orbitofrontal cortex
PFC	prefrontal cortex
pg	picogram
PSS	Perceived Stress Scale
RDoC	Research Domain Criteria
ROI	Region of interest
SASKO	Social Anxiety and Social Competence Deficits Scale
SCID	Structured clinical interview for DSM IV

SNc	pars compacta of the substantia nigra
SNR	signal-to-noise ratio
SPM	statistical parametric mapping software
STAI	State and Trait Anxiety Inventory
TE	echo time
TLFB	Alcohol Timeline Followback
TPJ	temporoparietal junction
TR	repetition time
VAS	visual analogue scale
VS	ventral striatum
VTA	ventral tegmantal area
%TWL	percentage total weight loss

1 INTRODUCTION

1.1 Mental Disorders – categorical versus dimensional nosology

Mental disorders are currently classified using standardized diagnostic systems that define symptom-based criteria for diagnostic categories, considering additional variables, such as the number, severity and duration of symptoms and disease episodes. The two most widely used diagnostic systems are the International Classification of Diseases in its 10th and 11th revision respectively (ICD10 and ICD11) (World Health Organization. Division of Mental, 1992, 2019) and the 5th revision of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) (Association, 2013). While the use of ICD10 and DSM-5 have enhanced the reliability and standardization of diagnosing mental disorders, current field studies investigating the ICD11 and DSM-5 reported limited reliability for several key diagnoses, such as Major Depressive Disorder (MDD, DSM-5) (Regier et al., 2013) and Dysthymic Disorder and Dysthymia respectively (ICD10 and ICD11). In addition, surveys of diagnosing mental health professionals indicated that only a little more than half of the respondents routinely went through diagnostic criteria to determine whether they apply to an individual patient and about 10% to 20% of health care professionals routinely used unspecified "residual" diagnostic categories, because they deemed that the clinical presentation did not conform to a specific diagnostic category. Moreover, most patients receive not one, but multiple diagnoses at the same time, while patients with the same diagnosis might not have a single symptom in common. Furthermore, the missing consideration of etiological factors and the simplification that results from applying categorical diagnoses or cut-offs to clinical symptomatology are increasingly questioned. It has been argued that this categorical approach cannot account for the clinical significance of symptoms that are below a specified cut-off (Helzer et al., 2006), but still cause clinically significant psychological strain and might disregard the shared neurobiological basis across different categorical diagnoses. This argument has been supported by findings show-

ing largely overlapping neural activation patterns across diagnostic categories, including schizophrenia, bipolar disorder, major depressive disorder, anxiety disorders, and obsessive-compulsive disorder (Sprooten et al., 2017). These and other observations have led to the establishment of initiatives that seek to put forward a dimensional understanding and nosology of mental disorders, e.g., the Research Domain Criteria (RDoC) initiative of the National Institute of Mental Health (Insel et al., 2010) or the Hierarchical Taxonomy Of Psychopathology (HiTOP) initiative (Kotov et al., 2017). The core aim of the RDoC initiative is to explore and identify central brain mechanisms and molecular entities that underlie the six postulated core behavioral domains or constructs: a negative valence system, a positive valence system, a cognitive system, a system for social processes, an arousal and regulatory processes system and a sensorimotor system. A central tool that has been used to investigate the neurocircuitry underlying the hypothesized domains and brain mechanisms, which are supposed to be shared across diagnostic categories, is functional magnetic resonance imaging (fMRI). While the RDoC initiative stimulated translational neuroscientific research, the efforts to identify the neurocircuitry underlying the RDoC constructs has been challenged by findings of the *Open Science Collaboration* that the rate of success replicating the findings of 100 neuroscientific experiments was as low as 40% (Collaboration, 2015) and recent meta-analytical results that indicate poor overall reliability across a range of 90 fMRI studies (Elliott et al., 2020). As a prominent example of this debate, Vul and colleagues in their publication "*Voodoo Correlations in Social Neuroscience*" argued that the reliability of fMRI measures places an upper limit to the strength of correlations between fMRI signal and behavioral measures (Vul et al., 2009). At the same time, effect size estimates derived from non-reliable measures can also be considered being prone to over-estimating the true population effects. While some of the arguments have been rebutted (Jabbi et al., 2011), the debate still seems important, since any effort to precisely define brain circuits and functions on an individual level using fMRI critically depends on the reliability and replicability of the applied fMRI measure.

In the following sections the general principles of magnetic resonance imaging and fMRI will be portrayed and the different approaches to determining the reliability of fMRI measures will be introduced, followed by a presentation of the current state of evidence on the reliability of task fMRI and an outline of the research questions of the doctoral thesis.

1.2 Magnetic resonance imaging (MRI)

Magnetic resonance imaging is a non-invasive in-vivo imaging technique that is based on the excitation of protons in a magnetic field by means of high-frequency pulses (Schneider, 2013; Weishaupt, 2014). Protons possess an intrinsic angular momentum, a so-called "*spin*". By applying a strong magnetic field, the spin axes of the protons can be aligned parallel or antiparallel to the magnetic field. This results in a precession motion of the protons proportional to the strength of the magnetic field with the so-called Lamor frequency. With the aid of high-frequency pulses switched perpendicular to the magnetic field and resonant with the Lamor frequency, the protons can be deflected out of the plane of the magnetic field in phase synchronism. This results in a measurable magnetic vector, the so-called transverse magnetization. After the high-frequency pulse is switched off, the protons return to their initial position. This results in longitudinal tilting of the protons back into the plane of the magnetic field. This is called "*T1 relaxation*" or "*longitudinal relaxation*". Furthermore, the in-phase spin of the protons slowly de-phases. This is called "*T2 relaxation*" or "*transverse relaxation*". During both relaxation processes, energy is emitted in the form of electromagnetic radiation, which can be detected and measured using a head coil. T1 and T2 relaxation are independent processes that occur in parallel. The duration of relaxation is measured as T1 and T2 relaxation time. The T1 relaxation time describes the duration until about 63% of the initial value of the longitudinal magnetization has been regained. The T2 relaxation time describes the duration until the transverse magnetization has decreased to approximately 37%. The MRI signal depends on the magnetic field strength as well as the T1 and T2 relaxation time and the proton density. The latter

three are thereby dependent on tissue composition, resulting in a different signal depending on the tissue. This allows different image contrasts to be calculated for imaging different tissues.

1.2.1 Functional magnetic resonance imaging (fMRI)

Functional magnetic resonance imaging (fMRI) aims to measure neuronal activity indirectly and noninvasively. It exploits the different magnetic properties of oxygenated and deoxygenated blood (Schneider, 2013; Weishaupt, 2014). Deoxygenated hemoglobin molecules are more magnetic and thus affect the relaxation behavior of protons in the environment. The resulting inhomogeneity in the magnetic field lead to a shorter transverse relaxation time (T2* relaxation) of the protons in the vicinity of deoxygenated hemoglobin molecules. As a result, a different signal can be measured depending on the amount of deoxygenated blood in a brain area. The Blood Oxygenated Level Dependent (BOLD) signal shows variability over time. Milliseconds after the onset of a stimulus, there is a decrease in the signal as an indication of an increase in the proportion of deoxygenated hemoglobin as a result of an increase in the oxygen demand of active neurons (Ernst & Hennig, 1994). This signal decrease is followed by an increase in the intensity of the BOLD signal due to an increase in the proportion of oxygenated hemoglobin above baseline level in terms of hemodynamic overcompensation of the initial reduced oxygen supply. The local adaptation of cerebral blood flow to the increasing oxygen demand is also called hemodynamic response. Studies have shown that during the processing of stimuli there is an initial reduction in the proportion of oxygenated blood. After approximately 1000 milliseconds, hemodynamic adaptation occurs and the proportion of oxygenated blood increases (Frostig et al., 1990). The maximum signal increase is reached 5 to 8 seconds after the onset of neuronal activity (Lee et al., 1995). Studies have demonstrated a relationship between neuronal activity and increase in BOLD signal (Logothetis et al., 2001). It has been shown that the level of neuronal activity is closely correlated with the level of the BOLD signal. However, the change in MR signal due to the BOLD effect is very small. Therefore, to map brain activity indirectly using fMRI, a stimulation condition (e.g.,

presentation of stimuli) is often compared with a control condition (e.g., resting) to form a contrast between a phase with pronounced brain activity and a phase with baseline activity. In addition, it must be taken into account that the measured signal is composed of the BOLD signal and noise. Noise in this context includes physiological effects (e.g., respiratory and body movements) and thermal factors (e.g., tissue heating) that may overlay the BOLD signal. The fMRI method is characterized by a relatively poor signal-to-noise ratio. However, the highest possible signal-to-noise ratio is important to detect effects of experimental conditions. Therefore, various measures are taken to reduce the noise. For this purpose, temporal and spatial adjustments or corrections of the data are made during the processing of the fMRI data, among other things.

1.3 Reliability of functional magnetic resonance imaging (fMRI)

Reliability in general can be considered as the consistency with which a measure or measurement produces similar results or values under consistent conditions. In the context of the doctoral thesis, which focusses on fMRI-based measures, specifically the test-retest reliability will be considered, which can be considered as the consistency with which repeated measurements that are gathered from a single rater who uses the same methods and testing conditions to produce similar results. In this case, the rater is considered to be the fMRI scanner and the associated methods.

First and foremost, the reliability of the fMRI BOLD signal in a voxel is influenced by the signal-to-noise ratio (SNR) in this voxel (Bennett & Miller, 2010). A high signal-to-noise ratio in a voxel would implicate that the primary source of variance within this voxel is due to "*true*" changes in regional cerebral blood flow and ultimately changes in neural activity, while a low SNR would implicate that the primary source of variance is due to noise or measurement errors.

Other parameters for expressing differences in signal strength between different brain areas or different time points are the image contrast-to-noise ratio (CNR) and the temporal CNR respectively. The latter can be considered as relative difference in

signal intensity in a specific voxel between e.g. two task conditions. This implies that a high temporal CNR is important, when trying to detect robust task condition effects.

Theoretically, the ability to measure reliable fMRI signal is limited by all factors that add error to the measurement of the signal. A detailed discussion of all factors is beyond the scope of the thesis, but specific factors that were identified by previous work should be outlined in brief.

1.3.1 Determinants of the fMRI Signal-to-Noise Ratio

Determinants of the fMRI SNR can be roughly divided into factors that relate to the properties of the fMRI scanner and measurement parameters and factors that relate to data processing and data analyses. The former includes the magnetic field strength of the fMRI scanner that is expressed in Tesla (T). Theoretically, an increase of the magnetic field strength by 100% would double the SNR. The practical increase in SNR by increasing the field strength from 1.5 to 3 T has been shown to be around 60-80%, which was attributed to counter-acting effects (e.g. increasing risk of susceptibility artefacts)(Hoenig et al., 2005). In addition, parameters of the fMRI acquisition were shown to have a substantial impact on the SNR and CNR. Specifically, higher voxel sizes (3mm vs. 1.5mm) were associated with a higher SNR. Other parameters that were shown to affect the SNR are the repetition time (TR), the echo time (TE), the bandwidth and the slice gap (Moser et al., 1996). However, the effect of each parameter might vary, depending on the field strength and the specific fMRI set-up, so that there is no generally optimal parameter set-up. Determinants of the SNR that relate to data processing and analyses are – among other – the spatial and temporal realignment procedures, temporal filtering of the EPI time series and spatial smoothing of the fMRI data (Bennett & Miller, 2010). All these parameters have been shown to significantly affect the SNR and CNR. In addition, changes to the parameters of the analytical pipeline have been shown to produce substantially different results and differences in the reliability of fMRI-based measures.

1.3.2 Estimates of fMRI reliability

There is a range of estimates of fMRI reliability, which assess different aspects of the reliability of the fMRI signal, but currently there is no common standard. While some estimates of reliability mainly focus on voxels that surpass a pre-defined statistical threshold, other reliability estimates focus on the magnitude of activity in specific brain voxels or across a set of voxels. While the former can be assessed using so-called cluster overlap methods, the latter can be assessed by computing the intraclass correlation (ICC) coefficients. Since these two methods were applied most frequently across previous studies assessing fMRI reliability, these are presented in detail. In addition, the concept of similarity estimation as an approach to measuring test-retest reliability is presented, because this approach tries to determine whether or not a specific subject can be identified by his or her neural activation patterns, which closely follows the idea of identifying neural biomarkers. It is acknowledged that a range of other reliability estimates exist that include similarity indices, receiver operating characteristic curves, Pearson correlation coefficients, Cohen's kappa index, predictive modeling and others. A detailed description of these approaches is beyond the scope of the submitted thesis.

Cluster overlap reliability estimates – the Jaccard and Dice coefficients

The cluster overlap method is used to assess the proportion of voxels in the brain that show significant activation above a pre-defined statistical threshold for both the test and retest session. The computed reliability estimates seek to represent the proportion of voxels that remain significant across repeated test sessions. Two approaches have been established to determine the cluster overlap. Both require a pre-defined statistical threshold for defining which voxels are to be considered significantly active. While there is no consistent standard in the literature, the most frequently used thresholds are an uncorrected threshold of $p < 0.001$ and $p < 0.01$ respectively. The first overlap method is the Jaccard coefficient. It is defined as the size of the intersection divided by the size of the union of the voxel sets that show significant activation

during the test session (A) and the retest session (B). The Jaccard coefficient can be computed as follows:

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Following the definition, the Jaccard coefficient can be interpreted as the number of overlapping voxels that show significant activation above a predefined statistical threshold during the test and retest sessions, divided by the number of all voxels that show significant activation during the test or retest session (Jaccard, 1902; Maitra, 2010). Consequently, a value of 1.0 reflects a 100% overlap between the voxel set showing significant activation during the test session with the voxel set showing significant activation during the retest session. In contrast to that, a value of 0.0 indicates that – in the case that there are any significant voxels – none of these overlap between the test and retest sessions.

The second overlap method is the Dice coefficient, which is defined as the number of super-threshold voxels that overlap between the test (A) and retest sessions (B) divided by the average number of significant voxels across sessions (A + B):

$$Dice(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

Following the definition, the Dice coefficient can be interpreted as the number of voxels that overlap between test and retest sessions divided by the number of significant voxels across both sessions. Much like the Jaccard coefficient, the values of the Dice coefficient range from 0 ("no overlap") to 1 ("perfect overlap").

It has to be noted that the results of the overlap methods are not independent from the statistical threshold that is used to define significant voxels (Duncan et al., 2009; Rombouts et al., 1998). With higher and more stringent statistical thresholds, the cluster overlap methods yields lower reliability estimates. This limitation should be considered when interpreting both coefficients. Furthermore, there is currently no consensus on cut-offs for "poor" and "good" ranges of overlap values (Bennett & Miller, 2010), which complicates the interpretation of the overlap statistics.

The Intraclass Correlation (ICC)

The ICC seeks to test whether the magnitude of activation in a specific voxel or in a set of voxels remains stable across test sessions. The ICC coefficient can be conceived as a correlation, which represents the correlation between the magnitudes of activation in two voxels or voxel sets. It has been argued that this measure might be more stringent than other fMRI reliability estimates, because it also requires near zero values to be stable over time (Bennett & Miller, 2010). In principle, six types of ICCs were described. It was suggested that the ICC(3,1)-type is most appropriate for the assessment of test-retest reliability of fMRI data (Ombao et al., 2016). Mathematically, this coefficient puts within-subject variance (σ^2_{within}) in relation to between-subject variance ($\sigma^2_{between}$) (Shrout & Fleiss, 1979) and is defined as:

$$ICC = \frac{(\sigma^2{}_{\text{between}} - \sigma^2{}_{\text{within}})}{(\sigma^2{}_{\text{between}} + \sigma^2{}_{\text{within}})}$$

Consequently, ICC values close to 1.0 represent near perfect agreement between the magnitudes of activation across test and retest sessions, while values close to 0.0 indicate that there is no agreement across sessions. According to Fleiss (1986), ICC coefficients < 0.4 can be regarded to represent poor reliability, while ICC coefficients >0.4 and <0.75 represent fair (< 0.6) to good (>0.6) reliability. According to the proposed classification, ICC coefficients that exceed a value of 0.75 represent good to excellent reliability (Fleiss, 1986). Like other reliability estimates, the ICC faces some limitations. Strictly speaking, the ICC coefficients are specific to the investigated data and sample, hampering generalization. In addition, the magnitude of the ICC depends on the between-subject variability, so that ICC estimates determined in samples with high inter-subject variance can yield higher ICC coefficients, even if the stability of the fMRI signal is similar.

Similarity

Another concept to represent test-retest of fMRI data are similarity analyses. This procedure seeks to capture the resemblance of brain activation patterns across time points, based on the alignment of high versus low brain activation values across the

brain (Frohner et al., 2019). In principle, this is achieved by computing all within- and between-subject correlations between brain activation patterns. The resulting coefficients are correlation coefficients that range from 'perfect' inverse relationship (-1.00) to a 'perfect' direct relationship (1.00) and can be represented in a matrix with a row and column for each measurement and session (Finn et al., 2015; Frohner et al., 2019). When the within-subject similarity estimate for a specific participant exceeds all between-subject similarity estimates, it can be assumed that the brain activation of this participant across the test and retest are more similar than the activation pattern of any other participant. In this case, authors proposed that a participant can be re-identified based on his or her neural activation pattern across sessions (Finn et al., 2015; Frohner et al., 2019).

1.4 Reliability of fMRI task measures – Relevance and Evidence

fMRI has been used since the 1990s to study individual differences in activation between fMRI task conditions. To this end, fMRI methods were optimized to produce robust activation during distinct task conditions in brain regions of interest. The robust within-subject effects and average group main effects led researchers to use the same fMRI tasks to study between-subject effects. The assumption underlying this approach is that if a brain region shows robust activation during a task condition, then inter-subject differences in the magnitude of this brain response pattern may underlie phenotypic differences in individual behavior. In a way, the focus of fMRI research partly moved from identifying average brain activation to studying differences in brain activation patterns between individuals. Applying fMRI to study between-subject effects heralds the potential to discovering biomarkers for brain disorders. Broadly speaking, a biomarker can be conceived as a biological indicator for disease risk, diagnosis or prognosis.

Stimulated by the RDoC initiative, a substantial amount of fMRI studies set out to determine the neurocircuitry underlying behaviors and mental disorders, and ultimately identify individual neural "biomarkers". This endeavor, however, critically depends on the test-retest reliability of the fMRI measure. On the one hand, the reliability of fMRI

determines the strength of possible correlations between fMRI measures and behavioral measures (Vul et al., 2009) and on the other hand poor reliability can lead to biased effect sizes estimates that would not mirror the true population effects size (Fried et al., 2017; Friedman, 1968; Wright, 2014). Consequently, fMRI reliability is critical for both research and clinical practice. Low reliability fMRI markers would be unsuitable as biomarkers and could not predict clinical outcomes.

Despite the undoubtedly high importance of reliability to the aims of most fMRI studies, only very few studies reported reliability estimates and even fewer studies have systematically analyzed fMRI reliability across studies. This has complicated individual studies on fMRI reliability with highly variable reported estimates, based on small test-retest samples and different methods to determine fMRI reliability, which has resulted in partially contradictory conclusions (Manuck et al., 2007; Nord et al., 2017).

The most comprehensive meta-analysis of task fMRI reliability was conducted by Elliott and colleagues (2020). Their meta-analysis included data from 90 experiments with N = 1008 subjects of 56 independent fMRI studies. Only 46% of the reported reliability scores that fell within the range of at least moderate reliability (ICC > 0.40), only 20% of the studies reported ICCs that fell in the range of good reliability (Elliott et al., 2020). The average ICC across all studies was 0.397 with a significant inter-study variability. The authors also conducted a moderator analysis, which indicated that neither task type, task design, task length, test-retest interval, ROI type (structural versus functional), nor sample type (healthy versus clinical) significantly affected the magnitude of the reliability estimates. Acknowledging the potential bias that might result from the heterogeneity between studies, the authors used data from the Human Connectome Project and the Dunedin study to assess the reliability of eleven task-fMRI measures, which were conducted on modern scanners using cutting-edge acquisition parameters, and standardized preprocessing pipelines. Nevertheless, results of reliability analyses of these data showed a poor reliability (ICC = 0.228) in a set of ROIs with no support for a better reliability, when ROIs targeted by the task were analyzed (mean ICC = .270 for target ROIs and .228 for non-target ROIs).

While the results might appear unexpected, it might point towards a fundamental problem because task-fMRI measures are based on contrasts between different task conditions. Due to the fact that the brain activation between task conditions frequently correlate substantially, a difference score between these two conditions eliminates substantial parts of the true variance, while the error variance is added. The authors concluded that difference scores will always have lower reliability than their constituent scores or conditions (Elliott et al., 2020). This, however, has implications for almost all fMRI studies that compute difference scores between task conditions. Surprisingly, the effect of the correlation between constituent task conditions on the reliability of the contrast between these conditions has been addressed by only a few studies and only for a limited selection of fMRI tasks. Two prototypical fMRI tasks that use a block design and compute a difference score between constituent task conditions are so-called cue-reactivity tasks and tasks that assess the self-concept.

1.4.1 The fMRI-based assessment of neural cue-reactivity

The investigation of neural responses to imagery with disease-specific context (e.g. pictures of alcohol) was used in the effort to determine the neurobiological underpinnings of the individual reactivity to such cues and associated behaviors. The so-called cue-reactivity paradigm are used to investigate a range of neurobiological processes including motivation, reward processing and – in the context of addictions – craving and incentive salience (Goldstein & Volkow, 2011; Hill-Bowen et al., 2020; Noori et al., 2016; Noori et al., 2012). The central assumption underlying the cue-reactivity tasks is stimuli that predicted availability or receipt of a reward (e.g. food, drug, sex) in the past can trigger arousal and changes in behavioral motivation and thereby elicit stimulus-associated responses such as an urge to seek food or drugs. These cue-induced responses or "cue-reactivity" can be physiological and behavioral. Tasks that assess neural cue-reactivity typically incorporate certain stimulus categories (drug, food, etc.) and stimulus modalities (e.g. visual cues, olfactory cues, auditory cues) or a combination thereof and present these in the framework of a block- or even-related-design to participants, while neural activation is measured using fMRI (see also section

1.2 for details on MRI methodology). Even though the procedure of cue-reactivity tasks can be considered as simple, studies reported that presentation of cues induced wide-spread brain activation in a wide network of brain areas involved in perception, attention, memory, reward and emotion processing (Hill-Bowen et al., 2020; Noori et al., 2016; Schulte et al., 2019). A recent meta-analysis compared the brain activation patterns across 196 cue-reactivity studies that incorporated visual drug or food cues (Hill-Bowen et al., 2020). The authors found convergent brain activation across all studies in brain areas including limbic, insular and frontal, parietal and occipital brain regions. Connectivity analyses of the fMRI task data indicated that neural cue-reactivity can be decomposed into elemental processes and indicated engagement of six functional subnetworks during cue-reactivity. Based on data from the Neurosynth database (Neurosynth.org), these six subnetworks were interpreted as visual perceptive network, visual association network, cognitive control network, salience network, valuation network and emotion processing network. These results suggest that multiple brain networks are engaged during cue-reactivity tasks, which contribute to cue-induced changes in perception and behavior.

While the so-called fMRI "cue-reactivity" paradigms were established for almost every stimulus category (e.g. alcohol, nicotine, cocaine, cannabis, heroin, amphetamine, food) and many stimulus modalities (e.g. visual cues, olfactory cues, auditory cues), reliability and temporal stability of the neural activation patterns during these tasks have not been thoroughly investigated yet. This seems surprising, as reliability is a prerequisite with many imaging studies aiming to capture individual neural "biomarkers" of brain function, and predicting future behavior based on these "biomarkers". This endeavor, however, seems to critically depend on the reliability of the specific fMRI task. Only a single study with n = 9 alcohol-dependent patients has 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) (Schacht et al., 2011). Results indicated good reliability in the right VS and DS, while poor reliability was reported for the left VS and DS. But even though many cue-reactivity studies also investigate brain activation beyond VS and DS, a comprehensive

analysis of the reliability of cue-reactivity patterns across the whole has not been conducted yet. With regards to food cues, only a single fMRI study assessed the test-retest reliability over a period of about 18 days in a limited range of a priori defined regions of interest (ROI) (bilateral insula, amygdala, orbitofrontal cortex, caudate and putamen) (R. Drew Sayer et al., 2016). Results of this study showed that only activation in the left orbitofrontal cortex ROI showed fair reliability, while brain activation in all other ROIs showed poor reliability. Overall, data on the reliability of the fMRI-based assessment of neural cue-reactivity is currently limited to a small range of ROIs and small samples, and a comprehensive analysis of whole-brain reliability is still missing.

1.4.2 The fMRI-based assessment of the self-concept

Empirical findings across substance use disorders and behavioral addictions support the importance of self-concept to onset and course of the disorder, and to changes in dysfunctional behavior (Corte & Zucker, 2008; Downey et al., 2000; Israelashvili et al., 2012; Leménager et al., 2018). Due to the fact that self-concept is conceived as stable construct, the investigation of the neurocircuitry of the self-concept has gained substantial interest in the effort to identify stable "biomarkers" for addictive disorders. fMRI studies have intended to address this question by the application of fMRI self-referential and fMRI self-recognition paradigms. During the so-called self-referential fMRI tasks, individuals are asked to judge their personality traits, physical appearance and personal preferences. This is then compared to the evaluation of another person (a close friend, a famous person, or a foreign person) (Fossati et al., 2004; Johnson et al., 2002; Lou et al., 2004). Neural activation during evaluation of the own person is then compared to activation elicited by evaluating the properties of other persons. The resulting neural activation patters are regarded as neurobiological substrate of different facets of the self-concept, such as the social and emotional self-concept.
Self-recognition paradigms are less explicit in the sense that individuals are not asked for their explicit judgement, but rather exposed by e.g. pictures of their own face or body and to pictures of others (Northoff et al., 2006). It was proposed that the neural activation, which is indicated by contrasting activation during presentation of pictures

of one-self vs. another person, reflects the neurobiological substrate of the physical self-concept.

Even though the fMRI-based assessment of aspects of the self-concept has gained significant attention in the last decades (Kim et al., 2018; Leménager et al., 2014; Lemenager et al., 2020), there is currently no data on the test-retest reliability of the fMRI tasks that are used to capture facets of the self-concept. Test-retest reliability of an fMRI task, however, is essential for capturing individual neural activation patters that reflect facets of the self-concept and for determining associations between these activation patterns and addictive behavior.

1.5 Research Questions and Hypotheses

Even though the advent of the RDoC initiative stimulated the conduct of a significant number of fMRI studies that aimed at identifying individual neural "biomarkers" for mental disorders, the test-retest reliability of the majority of the applied fMRI tasks remains largely unknown (Elliott et al., 2020). This, however, is an essential prerequisite for identifying individual neural "biomarkers", and for establishing associations between neural brain activation patterns and behavior. In addition, previous research highlighted the negative impact that might arise from computing difference scores between two correlated constituent task conditions, when applying fMRI tasks (Elliott et al., 2020; Infantolino et al., 2018). Even though there is a high probability that this also affects the results of fMRI cue-reactivity tasks and fMRI studies that investigate facets of the self-concept, this has not been thoroughly investigated yet. This, however, seems critical to determine, whether or not these fMRI-based measures fulfill the prerequisites to reliably assess neural brain activation, establish robust associations with behavior and estimate effect sizes. Against this background, the research questions and hypotheses were derived. Please note that parts of this thesis have been published (Study 1: (Bach et al., 2022) Study 2: (Bach, Grosshans, Koopmann, Kienle, et al., 2021), and Study 3: (Bach, Hill, et al., 2021)) by the doctoral candidate as first author. Therefore, certain text passages, tables and figures of this thesis will be

identical to these publications (the respective publisher has granted the respective rights for the use of this content in the framework of the doctoral thesis).

Research Question 1 (Study 1, 2 and 3): Is the test-retest reliability of the constituent conditions of an fMRI block design paradigm higher than the reliability of the difference score, which is created by subtracting brain activation during one condition from another condition and commonly used across a range of block design fMRI paradigms?

Hypothesis 1: The test-retest reliability of the constituting fMRI task conditions (e.g. alcohol and neutral), estimated by computing the ICC, Jaccard, and Dice coefficients exceeds the reliability of the difference score (e.g. alcohol-neutral) (Infantolino et al., 2018).

Research Question 2 (Study 1, 2 and 3): Does the level of correlation between the constituent conditions of an fMRI paradigm determine the level of the resulting reliability of the difference score? How does the test-retest reliability of the difference score of an fMRI food-cue reactivity paradigm compare to the test-retest reliability of validated psychometric scales used to examining subjective food craving?

Hypothesis 2A: A high correlation between the two constituting conditions results in a low reliability of the difference score.

Hypothesis 2B: The test-retest reliability of validated psychometric scales that assess food craving (study 2) and facets of the self-concept (study 3) is higher, compared to the reliability of the difference scores capturing food-cue induced brain activation ("food – neutral", study 2) and self-concept related processes ("self – familiar and unknown persons", study 3) respectively.

Research Question 3 (Study 1): Does the level of test-retest reliability of the difference score affect the ability to establish associations with external behavioral variables?

Hypothesis 3: The correlation between the difference score, contrasting two fMRI task conditions, and alcohol craving (external behavioral variable) is lower compared to the correlation between the constituent task conditions and the same external variable.

2 EMPIRICAL STUDIES

2.1 Study 1 - Test-retest reliability of fMRI-based assessments of alcohol cue-reactivity: is there light at the end of the MRI tube?[1]

2.1.1 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 two 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 vs. 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. While 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 vs. neutral" showed poor overall reliability (ICC<0.40), which was related to the inter-correlation 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 effects sizes. Future research on alcohol cue-reactivity should be cautioned by the low reliability of the common "alcohol vs. neutral" difference contrast. We propose that this limitation can be

[1] Bach P, Reinhard I, Koopmann A, Bumb JM, Sommer WH, Vollstädt-Klein S, Kiefer F. Test-retest reliability of neural alcohol cue-reactivity: Is there light at the end of the magnetic resonance imaging tube? Addict Biol. 2022 Jan;27(1):e13069. doi: 10.1111/adb.13069. Epub 2021 Jun 15. PMID: 34132011.

overcome by using the constituent task conditions as an individual difference measure, when intending to longitudinally monitor brain responses.

2.1.2 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 (Insel et al., 2010; Voon et al., 2020). 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 behaviors, such as craving and relapse (Bach et al., 2020; Grüsser et al., 2004; Karl Mann et al., 2014; Schacht et al., 2017). While 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) (Schacht et al., 2011). This seems surprising, because reliability is a prerequisite for the aim of many imaging studies, i.e. associating brain activation with behavioral variables (e.g. craving), predicting future behavior and developing treatment responsive biomarkers and neuroscience-based treatments for alcohol addiction (Heilig et al., 2019; Heilig et al., 2016).

A recent meta-analysis pointed out that the overall reliability of task fMRI across different tasks was poor (Elliott et al., 2020). 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 (Infantolino et al.,

2018; Peter et al., 1993). 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 different contrast alcohol vs. neutral that is commonly computed in alcohol cue-reactivity 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 to the difference contrast alcohol vs. neutral. Specifically, we hypothesized that the overlap of significantly activated voxels between first and second fMRI assessment, indexed by the Jaccard and Dice coefficients, would be higher for the neutral and alcohol contrast conditions, compared to 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 coefficient, for the neutral and alcohol contrast conditions, compared to 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 to the difference contrast.

2.1.3 Methods

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, 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 session. 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 three-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 program that runs about 21 days with a daily multiprofessional medically-supervised therapy schedule, here termed Intensified withdrawal treatment) (Loeber et al., 2009).

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 (5 to 9 per patients à 60 to 90 minutes each) between baseline and 2nd fMRI session. Patients received multiple sessions of cue-exposure treatment, i.e. they were exposed to their favorite 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, 50mg per 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 (Bach et al., 2020).

Subgroup 4: The fourth subgroup (IWT + CET II group) were n=29 patients that received about 9 CET sessions (à 60 to 90 minutes each) in addition to IWT. The details of the study are described in detail in a previous publication (Vollstadt-Klein et al., 2011).

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 (Kiefer et al., 2015).

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

Assessment

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

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 (Bach et al., 2020; Vollstadt-Klein et al., 2012). 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 5 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 (5 pictures per block) and 9 blocks of neutral pictures (5 pictures per block) were displayed. Each picture was presented for 4 seconds. 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 minutes 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 Picture Viewing Task) there was no rating phase and the task consisted of 12 blocks of the same alcohol stimuli and 12 blocks of the same neutral stimuli that were presented

for 4 seconds 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 minutes. 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).

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)(Frohner et al., 2019). Individual contrast images of the different task conditions served as input for the reliability analyses.

Intraclass correlation

Reliability of fMRI data was estimated for each voxel of the brain activation maps by computing the intraclass correlation (ICC) coefficients between time points. It was argued that the $ICC_{(3,1)}$-type is most appropriate for assessing longitudinal fMRI data (Ombao et al., 2016). Hence, the $ICC_{(3,1)}$-type was computed (for details see Supplementary Methods) for the difference contrast alcohol vs. 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 (Tzourio-Mazoyer et al., 2002). For details see Supplementary Methods.

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 first and second fMRI (see Supplementary Methods). Repeated measures analyses of variance models with the factors contrast category (alcohol, neutral, alcohol vs. neutral) and groups (5 patient subgroups) were used to test main effects and interactions on the magnitude of the Jaccard and Dice coefficients.

Similarity

We calculated the similarity of the fMRI activation maps using the fmreli toolbox (Frohner et al., 2019). 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 (Finn et al., 2015; Frohner et al., 2019).

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 vs. neutral difference contrast and the former two conditions using the fmreli toolbox.

Analyses of group-level fRMI activation

On a group level, imaging data were analyzed using full factorial models with the factor time (1st and 2nd 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}<.05, we determined a combined voxel-wise-[p<.001] and cluster-extent-threshold [k ≥ 110] by running 10.000 permutations by

Monte Carlo simulations (the estimated smoothness was x/y/z = 10.22/10.753/9.03 mm) using in the Neuroelf analysis package (www.neuroelf.net) (Bennett et al., 2009).

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 (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 inter-correlation 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 to the population parameter (Hunter & Schmidt, 1994), we assessed how test–retest reliability (or lack thereof) would attenuate small, medium, and large effect sizes resulting from clinical studies (see Supplementary Methods).

2.1.4 Results

Sample characteristics

The patient subgroups showed similar values across demographical and psychometric variables with the exception of the BDI scores and time of abstinence (see Table 2.1). 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 baseline and second assessment. Comparison between fMRI sessions showed a significant effect of time for the magnitude of AUQ ($F=15.20$, $p<0.001$) and OCDS scores ($F=60.61$, $p<0.001$) and a significant interaction between study subgroup and time for the OCDS ($F=8.69$, $p=0.049$), while 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 to the other study subgroups (post-hoc tests: $p < 0.044$, see Table 2.1).

Table 2.1 Demographic data, alcohol use and severity measures for patient groups with available imaging data for both time points (baseline and week two scan).

Subgroup	1	2	3	4	5		
	IWT (ALCUEPV) (n=21)	IWT + CET I (ALCUEPV) (n=42)	IWT + NTX (ALCUEPV) (n=20)	IWT + CET II (ALCUE) (n = 29)	IWT + CET + DCS (ALCUE) (n=32)	Statistics	Significance
Demographical variables							
Age (years)	47.6 (10.5)	46.6 (10.0)	48.3 (8.4)	47.5 (10.3)	44.0 (10.1)	$F_{(4,139)} = .801$	p=.526
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	Z = 12.745	p=.371
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)	$F_{(4,136)} = 2.138$	p=.079
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)	$F_{(4,135)} = 1.555$	p=.190
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)	$F_{(4,135)} = 1.516$	p=.201
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)	$F_{(4,135)} = 1.515$	p=.201
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)	$F_{(4,139)} = 8.095$	**p < .001***
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)	$F_{(4,139)} = 5.297$	**p = .001***
Days between 1st and 2nd fMRI session	16.1 (3.6)	16.9 (3.4)	15.6 (3.6)	16.2 (4.7)	18.2 (4.4)	$F_{(4,138)} = 1.846$	p = .123
Smoker (yes/no)	10:11	27:15	13:7	21:8	24:8	$Chi^2_{(4)} = 4.910$	p=.297
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)	$F_{(4,100)} = 0.968$	p=.428
1st fMRI - 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)	$F_{(4,139)} = 1.080$	p=.369
FTND (total score)	3.7 (3.8)	3.7 (3.3)	5.5 (2.7)	5.0 (3.2)	4.7 (2.9)	$F_{(4,114)} = 1.369$	p=.249
ADS (total score)	15.7 (6.3)	16.2 (6.7)	12.9 (5.9)	16.6 (5.7)	14.8 (7.4)	$F_{(4,139)} = 1.215$	p=.307
STAI (trait sumscore)	44.2 (15.9)	45.8 (12.0)	37.9 (10.8)	43.2 (9.6)	46.3 (10.6)	$F_{(4,135)} = 1.868$	p=.120

STAI (state sumscore)	40.2 (12.8)	43.5 (12.0)	36.9 (8.3)	43.4 (10.6)	45.0 (10.1)	$F_{(4,135)} = 2.012$	p=.096
AUQ (total score)	12.9 (5.3)	12.6 (5.8)	-	11.2 (4.9)	13.4 (7.4)	$F_{(3,111)} = .717$	p=.544
BDI (total score)	14.5 (11.7)	15.9 (10.1)°	9.7 (8.0)°	10.3 (7.4)	11.6 (8.9)	$F_{(4,138)} = 2.495$	p=.046*
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)	$F_{(4,124)} = 2.378$	p=.055
AUQ (total score)	10.3 (3.4)	9.9 (3.1)	-	9.3 (2.4)	10.9 (3.4)	$F_{(3,111)} = 1.317$	p=.273

ADS = Alcohol Dependence Scale; BDI = Beck Depression Inventory; FTND = Fagerstroem Test for Nicotine Dependence; OCDS = Obsessive-Compulsive Drinking Scale; STAI = State-Trait-Anxiety Inventory; SD = standard deviation; * = significant differences $p < 0.05$; ° = significant post-hoc test $p < 0.05$; IWT = subgroup receiving treatment as usual (i.e. multidisciplinary three-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 + NTX = subgroup receiving daily naltrexone (NTX) in addition to IWT; 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)

Robust main effects of the alcohol vs. 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 Supplementary Table S 2.1). There was no significant main effect of time or subgroup on the magnitude of alcohol cue-induced brain activation (contrast: alcohol vs. neutral).

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 (ICC_{Mean}>0.40, see Table 2.2) and good to excellent local reliability (ICC>0.75), mainly in the occipital gyri and to a lesser extent in the inferior and superior frontal gyri (see Figure 2.2) and moderate to good reliability (ICC>0.4) in large clusters of brain areas including mesolimbic brain regions (insula, putamen, caudate), as well as temporal and parietal gyri (see Supplementary Figure S1). The patterns of brain areas depicting moderate and good reliability for the neutral contrast condition resembled these findings.

Similarity analyses showed higher within-subject similarity compared to between-subject similarity over time for the alcohol and neutral contrast conditions ($t_{Alcohol}$>=11.02, p<0.001, $t_{Neutral}$>=9.85, p<0.001) with overall high within-subject similarity values ($r_{Alcohol}$=0.71, $r_{Neutral}$=0.69), which is visible in a prominent diagonal in the similarity matrices of the alcohol and neutral contrast maps (see Figure 2.3). 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 (seeSupplementary Table S 2.3). Across both statistical thresholds, both overlap indices were significantly higher for the alcohol and neutral contrast conditions than for the alcohol vs. neutral difference contrast ($F_{(2,278)}$>=10.717, p<0.001, for detailed results see Supplementary Results).

Table 2.2 Mean ICC values across groups and contrast conditions

A] Difference Contrast: alcohol-neutral

	Group	IWT (ALCU-EPV) (n=21)	IWT + CET II (ALCUE) (n = 29)	IWT + CET I (ALCUEPV) (n=42)	IWT + CET + DCS (AL-CUE) (n=32)	IWT + NTX (ALCUEPV) (n=20)
Group	**Mean (SD) [range]**	0.23 (.23) *[-.70 – .89]*	0.18 (.18) *[-.63 – .70]*	0.06 (.06) *[-.57 – .69]*	0.00 (.00) *[-.73 – .67]*	0.18 (.19) *[-.79 – .84]*
IWT (n=21)	0.23 (.23)		$t_{(43)} = 0.820$	**$t_{(61)} = 3.245$*** **	**$t_{(51)} = 3.775$*** **	$t_{(39)} = 0.931$
IWT + CET II (n = 29)	0.18 (.18)			**$t_{(69)} = 2.673$** **	**$t_{(59)} = 3.461$*** **	$t_{(47)} = 0.097$
IWT + CET I (n=42)	0.06 (.06)				$t_{(72)} = 1.377$	**$t_{(60)} = 2.433$***
IWT + CET+DCS (n=32)	0.00 (.00)					**$t_{(50)} = 3.103$** **

B] Contrast: alcohol

	Group	IWT (ALCU-EPV) (n=21)	IWT + CET II (ALCUE) (n = 29)	IWT + CET I (ALCUEPV) (n=42)	IWT + CET + DCS (AL-CUE) (n=32)	IWT + NTX (ALCUEPV) (n=20)
Group	**Mean (SD) [range]**	0.40 (.25) *[-.79 – .94]*	0.54 (.22) *[-.53 – .97]*	0.30 (.23) *[-.47 – .92]*	0.43 (.25) *[-.78 – .96]*	0.38 (.25) *[-.63 – .96]*
IWT (n=21)	0.40 (.25)		$t_{(43)} = 2.041$	$t_{(61)} = 1.661$	$t_{(51)} = 0.397$	$t_{(39)} = 0.239$
IWT + CET II (n = 29)	0.54 (.22)			**$t_{(69)} = 4.435$*** **	$t_{(59)} = 1.790$	**$t_{(47)} = 2.305$** **
IWT + CET I (n=42)	0.30 (.23)				**$t_{(72)} = 2.379$** **	$t_{(60)} = 1.353$
IWT + CET+DCS (n=32)	0.43 (.25)					$t_{(50)} = 0.656$

C] Contrast: neutral

	Group	IWT (ALCU-EPV) (n=21)	IWT + CET II (ALCUE) (n = 29)	IWT + CET I (ALCUEPV) (n=42)	IWT + CET + DCS (AL-CUE) (n=32)	IWT + NTX (ALCUEPV) (n=20)
Group	**Mean (SD) [range]**	0.25 (.29) *[-.72 – .95]*	0.46 (.22) *[-.49 – .96]*	0.27 (.24) *[-.58 – .92]*	0.21 (.31) *[-.73 – .96]*	0.39 (.24) *[-.66 – .96]*
IWT (n=21)	0.25 (.29)		**$t_{(43)}$ = 2.998*** **	$t_{(61)}$ = 0.312	$t_{(51)}$ = 0.396	$t_{(39)}$ = 1.787
IWT + CET II (n = 29)	0.46 (.22)			**$t_{(69)}$ = 3.474*** **	**$t_{(59)}$ = 3.569*** **	$t_{(47)}$ = 0.997
IWT + CET I (n=42)	0.27 (.24)				$t_{(72)}$ = 0.858	**$t_{(60)}$ = 2.017***
IWT + CET+DCS (n=32)	0.21 (.31)					**$t_{(50)}$ = 2.231***

* p < 0.05, ** p < 0.01, *** p < 0.005; IWT = subgroup receiving treatment as usual (i.e. multidisciplinary three-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; 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)

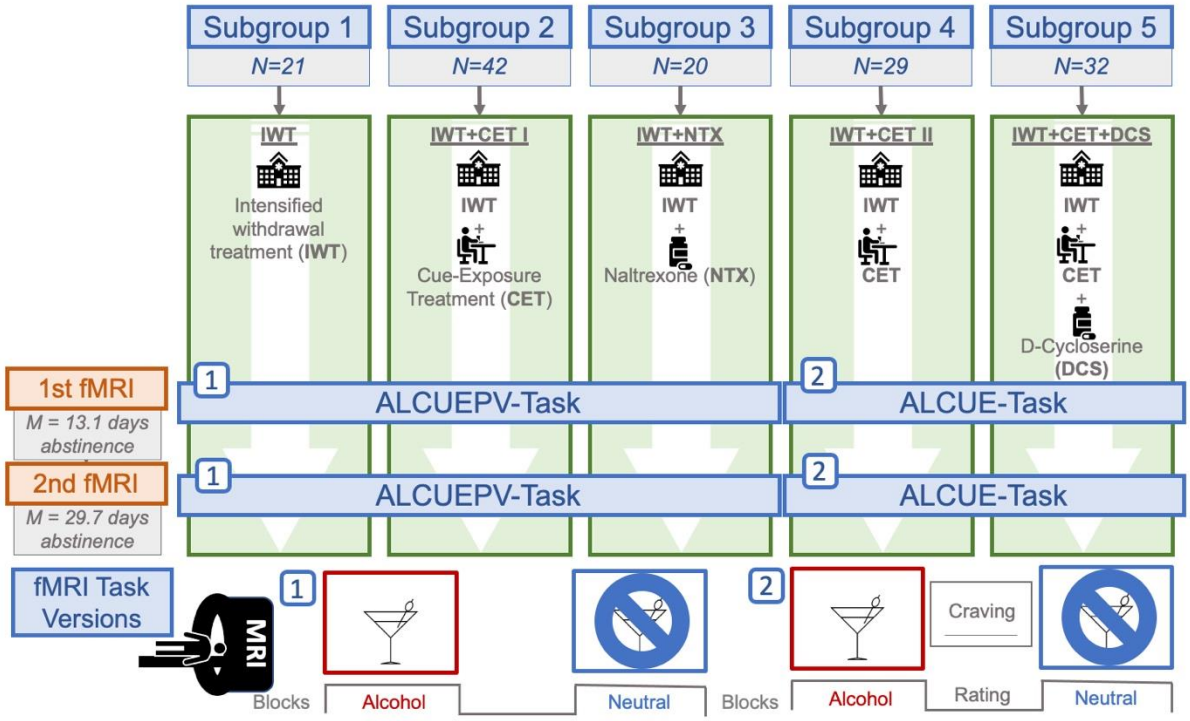
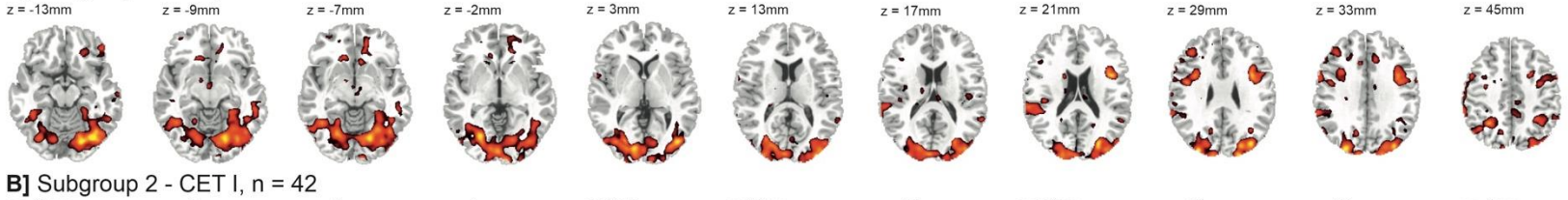


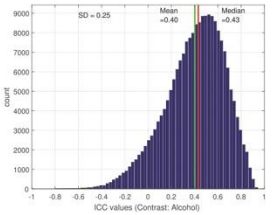
Figure 2.1 Depiction of the five different study subgroups. All patients received treatment as usual over two weeks (i.e. multidisciplinary intensified withdrawal treatment [IWT]). Subgroup 1 received no additional study treatment and hence serves as the reference group in the current analyses. Subgroups 2 and 4 received five to nine cue-exposure treatment (CET) sessions over two weeks in addition to IWT. Subgroup 3 received 50mg oral naltrexone (NTX) daily over two weeks in addition to IWT. Subgroup 5 received nine doses of 50-mg dose of d-cycloserine (DCS) concurrently with nine CET sessions over two 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 pseudo-randomized order. In contrast to the ALCUE task, the ALCUEPV task did not include a rating phase in-between successive picture blocks.

Brain regions with ICC values > 0.75 (“good“) for the task contrast „alcohol“

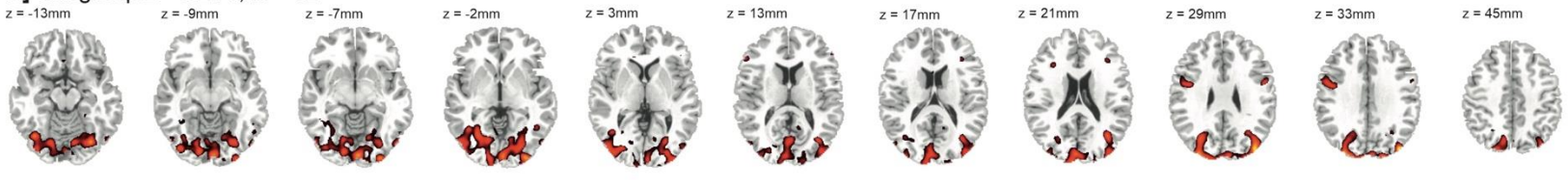
Histograms of ICC values

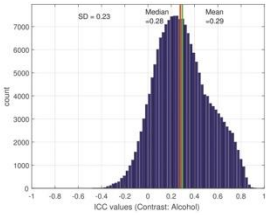
A] Subgroup 1 - IWT, n = 21



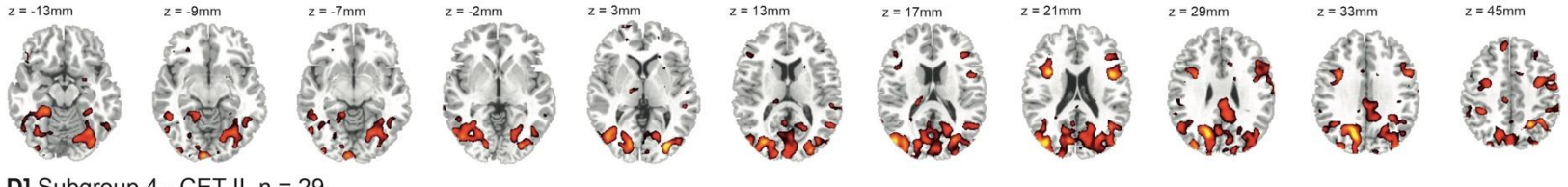


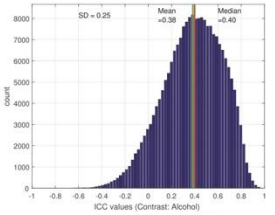
B] Subgroup 2 - CET I, n = 42



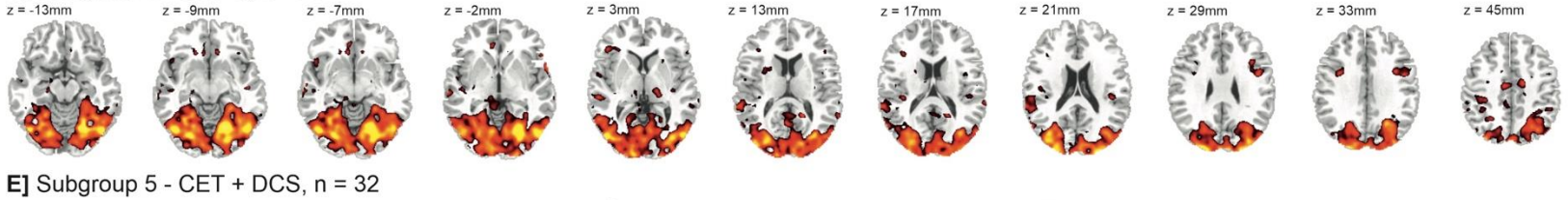


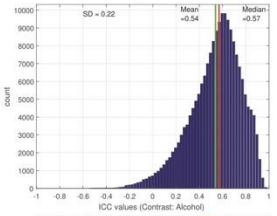
C] Subgroup 3 - IWT+NTX, n = 20



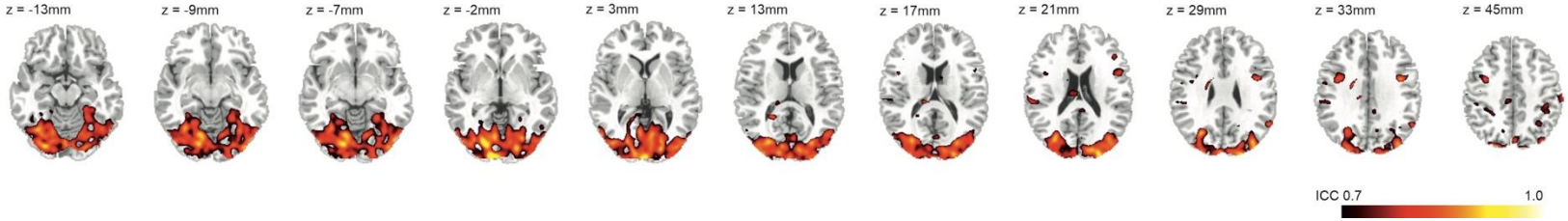


D] Subgroup 4 - CET II, n = 29





E] Subgroup 5 - CET + DCS, n = 32



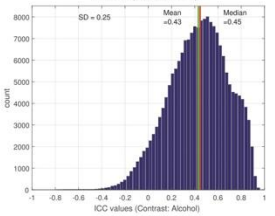


Figure 2.2 Depiction of brain areas that show good to excellent reliability or the alcohol condition contrast (Intraclass correlation [ICC] > 0.75) for the different study groups (A to E, left column) and histograms depicting the distribution of ICC values and the mean and median for the different study groups (right column).

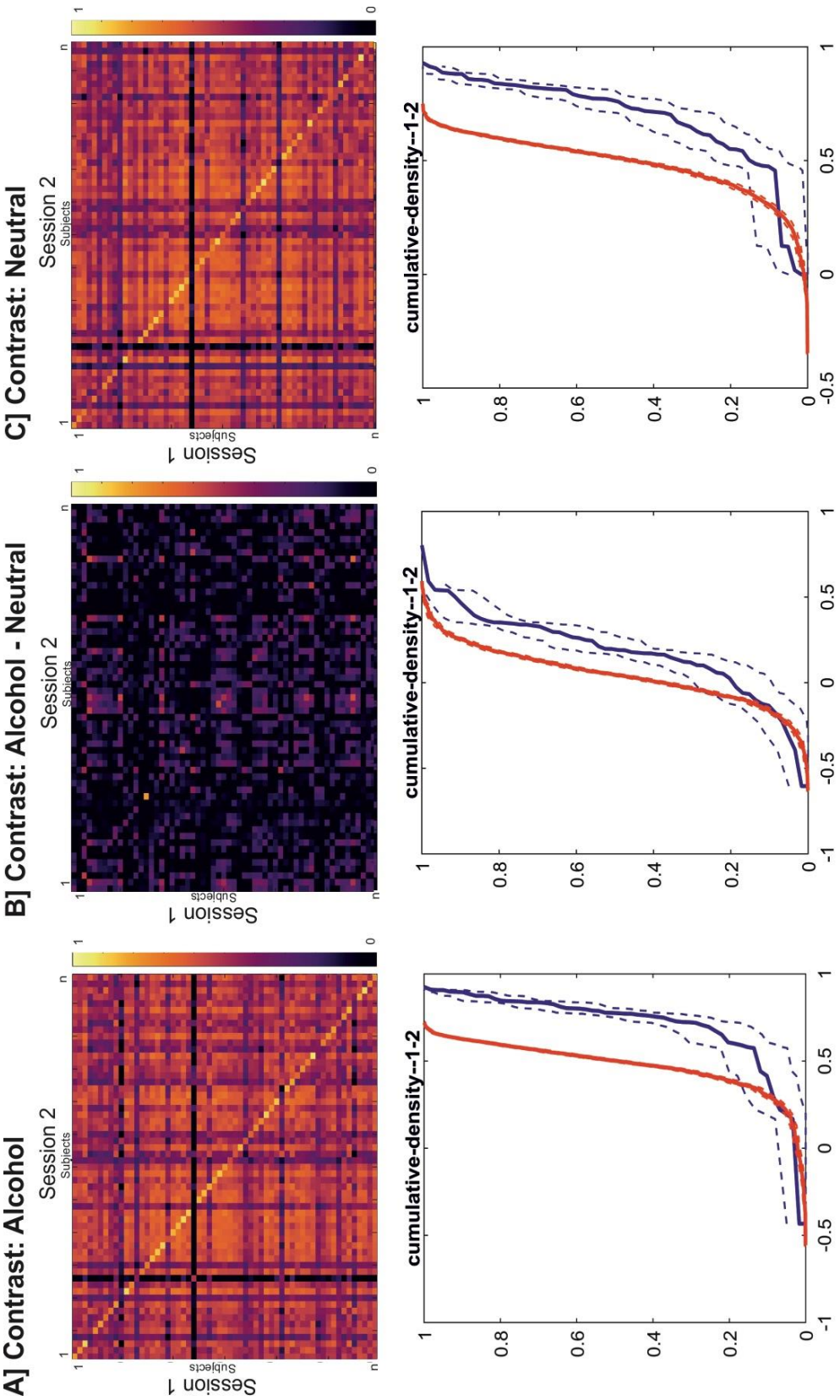


Figure 2.3 Similarity maps (upper row) and empirical cumulative distribution functions (lower row – red lines: between-subject similarity, blue lines: within-subject similarity) for longitudinal comparisons (1st and 2nd fMRI session) for the three contrasts: A] alcohol, B] alcohol-neutral and C] neutral.

The diagonal of each color matrix represents the within-subject similarity values. Re-identification of a subject based on the neural activation map is affirmed 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 B] the alcohol-neutral contrast.

Poor mean test-retest reliability of the of the alcohol vs. 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_{Mean}<0.4)(see Table 2.2). 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 2.4). Further analyses indicated that several additional brain areas showed moderate to good reliability, such as the occipital gyri, parts of the putamen and caudate, as well as parts of the inferior and superior frontal gyri (see Supplementary Figure S 2.1).

Similarity analyses showed that overall similarity values for the difference contrast were low ($r_{Alcohol\text{-}Neutral}$=0.19) and within-subject similarity did not exceed between subject similarity ($t_{Alcohol\text{-}Neutral}$<=1.83, p>0.05) (see Figure 2.3). This reflected in a low proportion of patients of 22% that could be re-identified based on their brain 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 vs. neutral difference contrast (see Supplementary Table S 2.3), compared to 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).

The atlas-based summary of mean ICC values for the difference contrast alcohol vs. 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 Supplementary Table S 2.2). 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 regards to the left and right dorsal and ventral striatum masks that were adapted from Schacht et al. (2011) to facilitate comparability, results show poor mean reliability (ICC<0.4) across all groups in all of the four masks (see Supplementary Table S 2.2).

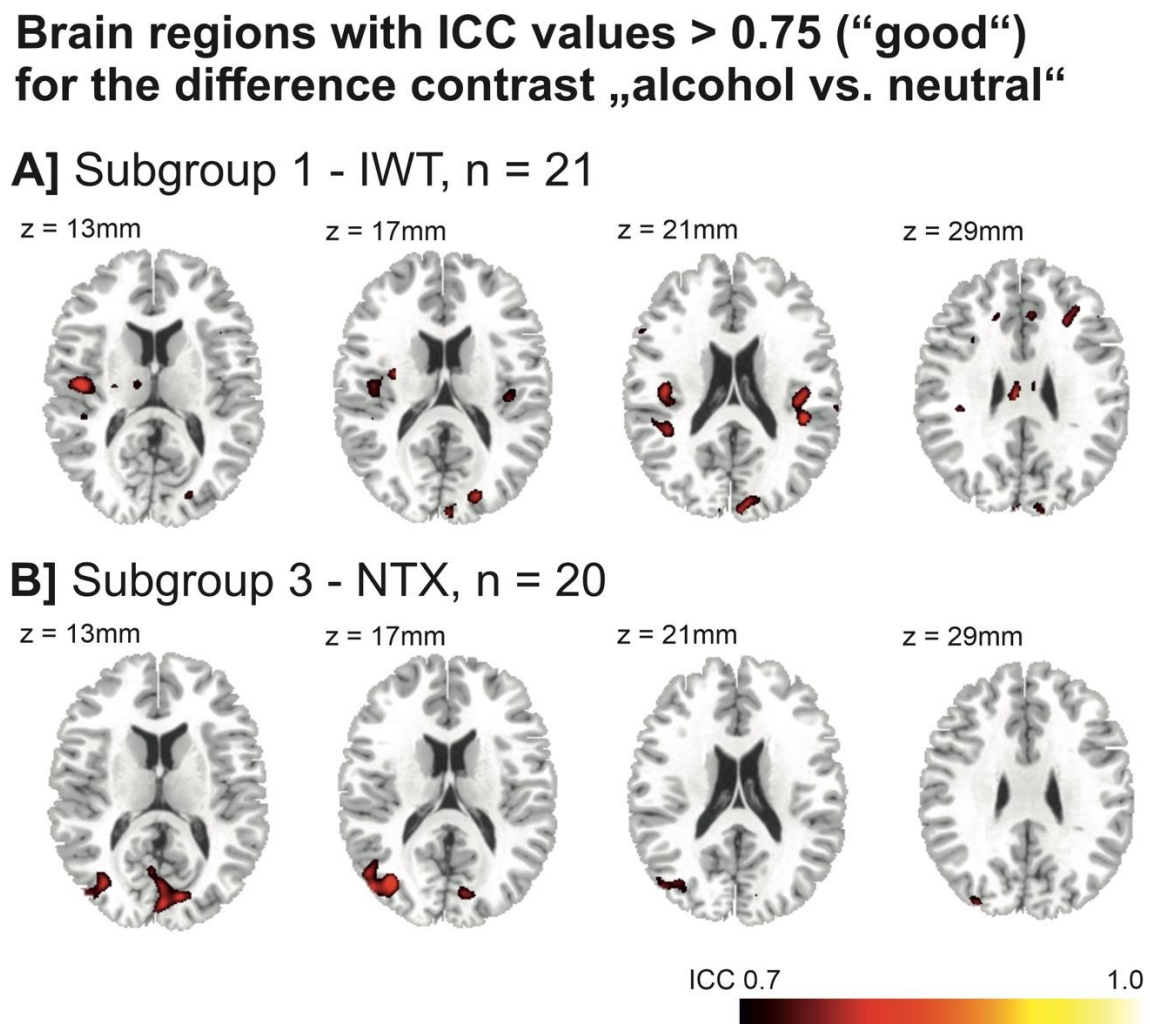


Figure 2.4 Depiction of brain areas that show good to excellent reliability for the difference contrast alcohol-neutral (Intraclass correlation [ICC] > 0.75) for two study groups (all other groups did not show ICCs > 0.75).

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).

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>=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 modeled according to the actual cue-reactivity and OCDS data, were always substantially lower, by about one third, compared to the correlation between the constituting conditions (V_1 and V_2) and the external variables (see Supplementary Table S 2.4 and Supplementary Figure S 2.3). 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 over-estimated the population effect sizes (see Supplementary Supplementary Figure S 2.4).

2.1.5 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) determines the capacity to establish associations with clinical data and estimate effects sizes from clinical trials.

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 session 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, 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 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.

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 inter-correlation 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 (2018). Their data on the difference contrast between face- and shape-matching trials of the well-established faces paradigm (Hariri et al., 2002; Infantolino et al., 2018) 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 (Elliott et al., 2020). A subsequent moderator analysis indicated that neither task type, task design, task length, test-retest interval, ROI type (i.e., structural versus functional), nor sample type (i.e., healthy versus clinical) signif-

icantly moderated reliability scores. While this seems unexpected, it might point towards a fundamental problem, i.e. 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. Elliot and colleagues (2019) concluded that difference scores will always have lower reliability than their constituent scores or conditions (Elliott et al., 2020). 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 heterogenous but locally satisfactory reliability to monitor treatment effects and treatment efficacy.

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 inter-correlation 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 to the correlation with the constituent task conditions (i.e. alcohol, 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 inter-correlation 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).

Overcoming low reliability of the alcohol vs. 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 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 (Infantolino et al., 2018). 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. cue-exposure treatment 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 regions of interest 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 measure for the stability of the cue-induced fMRI signal.

We applied this approach to our dataset, i.e. we specified the putamen and caudate as ROIs based on reliability scores and their roles for the cognitive process under investigation, i.e. alcohol craving (Noori et al., 2016). 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 inter-correlation of the task conditions for establishing associations with clinical data.

Limitations

It could be argued that the inclusion of patients without any treatment might be favorable with regards 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 version 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 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 leveled 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.

2.1.6 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 sizes 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.

2.1.7 Supplements

Supplementary Material

Detailed description of exclusion and inclusion criteria

Patients were recruited from the inpatient unit of the Department of Addictive Behaviour and Addiction Medicine at the Central Institute of Mental Health (Mannheim, Germany) from 2010 to 2017. All participants were required to be aged between 18 to 75 years. Patients also had to meet all following inclusion criteria: i) diagnosis of an alcohol-dependence according to the Diagnostic Statistical Manual of Mental Disorders (DSM-IV), ii) right handedness, and iii) completed detoxification (i.e. treatment - if necessary - of withdrawal symptoms with short-acting benzodiazepines had to be completed for at least 5x elimination half-life times). Patients were excluded if they met any of the following exclusion criteria: i) comorbid axis-I disorder (other than nicotine-dependence, assessed using the standardized clinical interview for DSM-IV axis I disorders, SKID i) in the last year, ii) treatment with psychotropic or anticonvulsive medications, iii) severe neurological or physiological disease [such as, but not limited

to stroke, aneurysm, dementia, epilepsy, liver cirrhosis], iv) positive drug urine screening on the day of testing, or v) contraindications for MRI scanning (i.e. pace-makers, metal implants, tattoos). On the first assessment day, all participants completed questionnaires, including the Beck Depression Inventory (BDI (Hautzinger, 2009), the Alcohol Dependence Scale (ADS (Kivlahan et al., 1989)), the Alcohol Urge Questionnaire (AUQ (Bohn et al., 1995)), the Obsessive Compulsive Drinking Scale (OCDS (Anton et al., 1995)) the Fagerstroem Test for Nicotine Dependence (FTND (Heatherton et al., 1991)) and the State Trait Anxiety Inventory (STAI; (Spielberger, 1983)). Substance use during the 90 days before the experiment was assessed using a semi-structured interview (Timeline Followback (Sobell et al., 1996)).

fMRI acquisition and pre-processing details

Functional activation during both alcohol cue-reactivity tasks was measured using a Siemens MAGNETOM 3 Tesla whole-body-tomograph (MAGNETOM Trio, TIM technology, Siemens, Erlangen, Germany). During the task, a total of 305 (ALCUE) and 303 (ALCUEPV) T2*-weighted echo-planar images (EPI) were acquired per participant using standardized imaging parameters (TR = 2.41 s, TE = 25 ms, flip angle = 80°, 42 slices, slice thickness = 2 mm, 1-mm gap, voxel dimensions 3 x 3 x 3 mm^3, FOV = 192 x 192 mm^2, 64 x 64 in-plane resolution). The first five T2*-weighted EPI images were eliminated from the datasets, in order to avoid bias due to magnetic saturation effects.

The functional EPI data were pre-processed according to standard procedures implemented in the statistical parametric mapping software for Matlab (SPM, Wellcome Department of Cognitive Neurology, London, UK) version 8 (i.e. spatial realignment, movement correction, normalization to a standard MNI [Montreal Neurological Institute, Quebec, Canada] EPI template and smoothing using an isotropic Gaussian kernel for group analysis [8 mm Full Width at Half Maximum]). Standardized quality checks were implemented for all datasets. Data were excluded if the spatial realignment or movement correction indicated excessive motion (>3 degrees of rotation or >3mm movement in any axis) or if visual inspection indicated poor fitting to the standard EPI

template. First-level statistics were computed for each participant, modelling the different experimental conditions (ALCUE: alcohol pictures, neutral pictures, rating phase or ALCUEPV: alcohol pictures, neutral pictures) in a generalized linear model including motion parameters as covariates. The contrast images for the alcohol and neutral condition, as well as the difference contrast between both conditions (alcohol-neutral) were computed. Due to the fact that alcohol cue-reactivity studies focussed on the difference contrast (alcohol – neutral), reliability analyses were conducted for this contrast. On the other hand, previous studies suggested that difference measures suffer from low inherent reliability, when the constituting conditions are correlated (Infantolino et al., 2018). The reason for this being that the variance of a difference score (e.g. alcohol-neutral) reflects the sum of the unique variance of both the neutral and alcohol condition, as well as their measurement error. Hence, any variance of the true scores that is shared between conditions is eliminated through subtraction, while condition-related error summates. This applies to all scenarios where the single conditions under investigation show a substantial inter-correlation, resulting in considerable shared variance of the true scores (Chiou & Spreng, 1996). Hence, we also estimated reliability separately for the alcohol and neutral conditions.

Reliability Analyses

Intraclass correlation coefficient

Reliability of fMRI data was estimated for each voxel of the brain activation maps by computing the intraclass correlation (ICC) coefficients between time points. According to Fleiss (1986), ICC coefficients lower than 0.4 represent poor reliability, ICCs between 0.4 and 0.75 represent fair (< 0.6) to good (>0.6) reliability, and ICCs higher than 0.75 represent good to excellent reliability (Fleiss, 1986). This coefficient sets within-subject variance (σ^2_{within}) in relation to between-subject variance ($\sigma^2_{between}$).:

$$ICC = \frac{(\sigma^2{}_{\text{between}} - \sigma^2{}_{\text{within}})}{(\sigma^2{}_{\text{between}} + \sigma^2{}_{\text{within}})}$$

ICC values were computed for the contrasts "alcohol-neutral", "neutral" and "alcohol" and for every study subgroup (1 to 5). We generated thresholded ICC brain maps, to

identify brain areas that show moderate to good (ICC > 0.4) and good to excellent (ICC > 0.75) reliability and we computed additional atlas-based mean ICC values for a standard set of anatomical brain regions (see below).

Atlas- and ROI-based summary measures

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 (Tzourio-Mazoyer et al., 2002). Additionally, we extracted the ICC values from the dorsal and ventral striatum using the ROI definition that was used by Schacht and colleagues (2011) in their analyses. All ICC values were extracted from the ROIs using the data extraction routine of the MarsBar software package (http://marsbar.sourceforge.net/) and exported into the IBM SPSS statistics software (version 25.0) or further analyses.

Jaccard and Dice Coefficients

Reliability of patterns of significant voxels was assessed using the modified Jaccard coefficient as an established and commonly used measure in fMRI reliability studies. It can be interpreted as the percentage of overlapping significant voxels above a predefined threshold (e.g. p < 0.001) within all significant voxels and is defined as the size of the intersection divided by the size of the union of the voxel sets (Jaccard, 1902; Maitra, 2010).

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Another measure of overlap between super-threshold voxels is the Dice coefficient. It was introduced to assess fMRI cluster overlap and has become an established measure of fMRI data reliability (Rombouts et al., 1997). It is calculated as the number of super-threshold voxels that overlap between sessions divided by the average number of significant voxels across sessions:

$$Dice(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

Both, Jaccard and Dice coefficients range from no overlap (0) to perfect overlap (1) between super-threshold voxels, but there is currently no consensus on specific values or cut-offs that would differentiate between "poor" and "good" values (Bennett & Miller, 2010). The interpretation is further complicated by the fact that cluster overlap methods depend on the statistical threshold used to define what is "active" and studies reported that the reliability of the cluster overlap method decreases, when the significance threshold is increased (Duncan et al., 2009; Rombouts et al., 1998). In the current analyses, we therefore used two common thresholds and compared the resulting indices ($p < 0.001$ and $p < 0.01$). Jaccard and Dice coefficients were determined for every patient comparing the baseline and 2nd fMRI for the different contrast images (alcohol-neutral; alcohol; neutral). Jaccard and Dice coefficients were exported into the IBM SPSS statistics software (version 25.0) and effects of contrast conditions and patient subgroups (n=5) were tested using a repeated measures analysis of variance model with contrast condition as within-subject factor and patient group as between-subject factor.

Supplementary Material – Analyses of associations between alcohol cue-reactivity with clinical data

We investigated associations between subjective craving, using the Obsessive Compulsive Drinking Scale (OCDS), and brain activation in the two regions of interest (putamen and caudate, which showed moderate to good reliability) during the neutral, alcohol and difference (alcohol-neutral) contrasts. The regions were defined by using the standardized anatomical masks as implemented in the automated anatomical labelling atlas (aal) atlas (Tzourio-Mazoyer et al., 2002) for the left and right caudate and putamen. Data on mean brain activation during the different contrast conditions ("alcohol", neutral", "alcohol-neutral") were extracted from the regions of interest using the data extraction routine of the MarsBar software package (http://marsbar.sourceforge.net/) and exported into the IBM SPSS statistics software (version 25.0) for further analyses.

Simulation study

We hypothesized that associations between the difference contrast and clinical variables are limited by the magnitude of the inter-correlation between the constituting contrast conditions. To fortify this assumption, we conducted a simulation study. 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. Specifically, we simulated and tested associations between the constituting contrast conditions (V_1 and V_2) and external variables (modeled after the OCDS scores) and between the difference contrast (V_1-V_2). In doing this, samples of size N=144 were generated under the following assumptions: the constituting contrast conditions alcohol (V_1) and neutral (V_2) are correlated with r=0.57, with means M_1=0.1035 and M_2=0.2272 as well as standard deviations S_1=0.3906 and S_2=0.8686 (analogue to the characteristics in our sample). For varying parameter values, correlations of external variables V_3 to V_{10} with V_1 and V_2 respectively were defined to range from r=0.1 to r=0.8, while the mean and standard deviation of V_3 to V_{10} was set to 15.8264 and 6.6557 respectively (analogue the characteristics of the OCDS values). We conducted 10.000 replications for each scenario V_3 to V_{10}.

Additionally, we assessed how test–retest reliability (or lack thereof) would attenuate small, medium, and large effect sizes resulting from clinical studies, following previously applied methodology (Fried et al., 2017; Friedman, 1968; Wright, 2014). We computed the reliability-adjusted effects sizes estimates, expressed as Cohen's d, which is an established effect size estimate in clinical studies. Adjusted Cohen's d values were determined for values ranging from 0.001 to 0.900 in steps of 0.001 for ICC values of 0.3 = poor reliability, 0.4 = cut-off between poor and moderate reliability, 0.6 = cur-off between moderate and good reliability, 0.75 = cut-off between good and excellent reliability and 0.9 = excellent reliability. In a first step Cohen's d were transformed to idealized r values according to Cohen (Cohen, 2013) using the following formula:

$$r_{idealized} = \frac{d}{\sqrt{d^2 + 4}}$$

In a second step, the r values were adjusted for reliability using the ICC values (Wright, 2014):

$$r_{adjusted} = r^2_{idealized} \times \sqrt{ICC}$$

In a last step, the adjusted r value was converted back to an adjusted Cohen's d using the formula (Friedman, 1968):

$$d_{adjusted} = \frac{2 \times r_{adjusted}}{\sqrt{1 - r^2_{adjusted}}}$$

Supplementary Results

Jaccard coefficient

Descriptive analyses of the Jaccard coefficients showed a large inter-individual variability across the contrast image conditions and across both statistical thresholds that were used to define significant voxels (see supplementary Table S3A-B). When the statistical threshold for defining super-threshold voxels was set to p<0.001, repeated measures analyses showed a significant main effect of contrast image category (neutral, alcohol and alcohol-neutral) ($F_{(2,278)}$=860.093, p<0.001) on the magnitude of the Jaccard indices, but no significant main effect of patient sub-group ($F_{(4,139)}$=0.954 p=0.435) or interaction between contrast category and patient sub-group ($F_{(8,278)}$=1.452 p = 0.175), such that the Jaccard coefficient were significantly higher for the neutral and alcohol condition contrast, compared to the difference contrast (alcohol-neutral) (both p<0.001, see supplementary Table S3A-B). For analyses where the statistical threshold for defining super-threshold voxels was set to p<0.01, results replicate the significant main effect of contrast image category ($F_{(2,278)}$=648.951, p<0.001), while neither the main effect of patient group $F_{(8,278)}$=0.945, p=0.480), nor the interaction ($F_{(4,139)}$=0.796, p=0.530) between both yielded significance.

Dice Coefficient

Across all contrast image conditions and across both statistical thresholds (p<0.001, p<0.01), descriptive analyses of the Dice coefficients showed a large inter-individual

variability (see supplementary Table S3C and 3D). For the statistical threshold of p<0.001, analyses showed a significant main effect of contrast image category (neutral, alcohol and alcohol-neutral) ($F_{(2,278)}$=948.150, p<0.001) on the magnitude of the Dice coefficient and a significant interaction effect between contrast category and patient sub-group ($F_{(8,278)}$=2.022, p=0.044), but no main effect of patient sub-group ($F_{(4,139)}$=0.985, p=0.418), such that the Dice coefficients were significantly higher for the neutral and alcohol condition contrast, compared to the difference contrast (both p<0.001, see supplementary Table S3C and 3D). For analyses where the statistical threshold for defining super-threshold voxels was set to p<0.01, results replicate the significant main effect of contrast image category ($F_{(2,278)}$=10.717 p < 0.001), while neither the main effect of patient group ($F_{(4,139)}$=1.409 p=0.234), not the interaction ($F_{(8,278)}$=1.376 p=0.207) between condition and group were significant.

Supplementary Tables

Supplementary Table S 2.1 Brain areas depicting significantly higher activation during alcohol picture blocks compared to neutral picture blocks during the 1st and 2nd fMRI session in the five patient subgroups (contrast: alcohol-neutral, combined voxel-wise- [$p < .001$] and cluster-extent-threshold [$k > 110$ voxel], corresponding to $p_{FWE} < .05$).

Side	Lobe	Brain areas	Cluster size (voxel)	MNI coordinates (x, y, z)			t_{max}
a) Group 1 (IWT)	n = 21						
T1							
L	Occipital, Temporal	Inferior Occipital Gyrus, Middle Occipital Gyrus, Fusiform Gyrus, Cerebellum, Lingual Gyrus, Calcarine, Cerebellum, Inferior Temporal Gyrus, Superior Occipital Gyrus	1714	-20	-94	-8	7.28
R	Occipital	Inferior Occipital Gyrus, Lingual Gyrus, Calcarine, Cerebellum, Middle Occipital Gyrus, Superior Occipital Gyrus, Cuneus, Fusiform	1116	18	-90	-22	5.59
L, R		Posterior Cingulum, Middle Cingulum, Precuneus, Posterior Cingulum, Precuneus, Middle Cingulum	1129	10	-38	26	5.39
T2							
L	Occipital, Temporal	Fusiform Gyrus, Inferior Occipital Gyrus, Middle Occipital Gyrus, Lingual Gyrus, Cerebellum,	2367	-20	-94	-10	8.54

		Cerebellum, Calcarine, Inferior Temporal Gyrus					
R	Occipital	Inferior Occipital Gyrus, Lingual Gyrus, Middle Occipital Gyrus, Calcarine, Fusiform Gyrus, Cerebellum, Superior Occipital Gyrus, Cerebellum	1022	34	-92	-2	6.30
L, R		Cerebellum	135	8	-74	-26	4.38
L, R		Middle Cingulum, Posterior Cingulum, Posterior Cingulum	100	8	-34	30	4.32
R	Parietal	Superior Parietal Gyrus, Angular Gyrus,	100	32	-68	52	4.10
b) Group 2 (CET I)	n = 29						
T1							
L, R	Occipital, Temporal	Middle Occipital Gyrus, Inferior Occipital Gyrus, Fusiform Gyrus, Lingual Gyrus, Cerebellum, Superior Occipital Gyrus, Lingual Gyrus, Inferior Temporal Gyrus	2705	-20	-98	-6	7.85
L, R	Occipital	Inferior Occipital Gyrus, Fusiform Gyrus, Lingual Gyrus, Cerebellum, Calcarine, Cerebellum, Middle Occipital Gyrus, Cuneus, Superior Occipital Gyrus, Calcarine	2024	24	-84	-12	6.79
T2							
L	Occipital	Middle Occipital Gyrus, Inferior Occipital Gyrus, Fusiform Gyrus, Lingual Gyrus, Calcarine, Cerebellum, Superior Occipital	2013	-22	-100	-6	8.59

		Gyrus, Cerebellum					
R	Occipital	Inferior Occipital Gyrus, Fusiform Gyrus, Lingual Gyrus, Middle Occipital Gyrus, Calcarine, Superior Occipital Gyrus, Cerebellum, Cuneus, Cerebellum	1611	28	-94	-8	7.85
c) Group 3 (CET II)	n = 42						
T1							
L, R	Occipital, Temporal	Fusiform Gyrus, Inferior Occipital Gyrus, Middle Occipital Gyrus, Inferior Occipital Gyrus, Lingual Gyrus, Fusiform Gyrus, Lingual Gyrus, Calcarine, Cerebellum, Cerebellum, Calcarine, Cerebellum, Middle Occipital Gyrus, Cerebellum, Superior Occipital Gyrus, Inferior Temporal Gyrus, Cuneus	4899	-26	-90	-12	9.25
L	Occipital, Temporal, Parietal	Angular Gyrus, Inferior Parietal Gyrus, Superior Parietal Gyrus, Middle Occipital Gyrus, Superior Occipital Gyrus, Middle Temporal Gyrus	693	-32	-70	52	5.45
L, R		Precuneus, Posterior Cingulum, Cuneus, Posterior Cingulum, Precuneus, Middle Cingulum, Calcarine	721	0	-36	26	5.06
L	Frontal	Precentral Gyrus, Middle Frontal Gyrus, Postcentral Gyrus	141	-48	-2	54	4.53
R	Occipital, Parietal	Superior Occipital Gyrus, Angular Gyrus, Middle Occipital Gyrus, Superior Parietal Gyrus	195	32	-66	40	4.09

T2							
L	Occipital, Temporal	Middle Occipital Gyrus, Fusiform Gyrus, Inferior Occipital Gyrus, Lingual Gyrus, Cerebellum, Cerebellum, Calcarine, Inferior Temporal Gyrus, Superior Occipital Gyrus	3285	-28	-90	-8	11.17
R	Occipital	Inferior Occipital Gyrus, Lingual Gyrus, Fusiform Gyrus, Cerebellum, Calcarine, Middle Occipital Gyrus, Superior Occipital Gyrus, Cuneus	2108	32	-94	-10	9.35
L	Occipital, Parietal	Superior Parietal Gyrus, Inferior Parietal Gyrus, Middle Occipital Gyrus, Superior Occipital Gyrus, Angular	663	-22	-62	42	5.47
R	Occipital, Parietal	Superior Occipital Gyrus, Angular Gyrus, Superior Parietal Gyrus, Middle Occipital Gyrus, Cuneus	680	30	-64	44	5.31
L	Frontal	Precentral Gyrus, Inferior Frontal Gyrus, Postcentral Gyrus,	581	-44	-8	44	5.00
L, R		Precuneus, Posterior Cingulum, Precuneus, Calcarine, Posterior Cingulum	354	-4	-54	26	4.45
d) Group 4 (CET + DCS)	n = 32						
T1							
L, R	Occipital, Temporal	Middle Occipital Gyrus, Fusiform Gyrus, Lingual Gyrus, Inferior Occipital Gyrus, Inferior Occipital	6430	-24	-100	-4	9.65

		Gyrus, Lingual Gyrus, Fusiform Gyrus, Calcarine, Middle Occipital Gyrus, Cerebellum, Calcarine, Cerebellum, Cerebellum, Cerebellum, Superior Occipital Gyrus, Superior Occipital Gyrus, Cuneus, Inferior Temporal Gyrus, Cuneus					
T2							
L, R	Occipital, Temporal	Middle Occipital Gyrus, Fusiform Gyrus, Lingual Gyrus, Inferior Occipital Gyrus, Inferior Occipital Gyrus, Lingual Gyrus, Fusiform Gyrus, Calcarine, Calcarine, Middle Occipital Gyrus, Cerebellum, Cerebellum, Cerebellum, Superior Occipital Gyrus, Superior Occipital Gyrus, Cerebellum, Cuneus, Cuneus, Inferior Temporal Gyrus, Inferior Temporal Gyrus	5811	-24	-98	-4	9.42
L, R		Precuneus, Posterior Cingulum	111	-2	-54	22	4.44
e) Group 5 (NTX)	n = 20						
T1							
L	Occipital	Inferior Occipital Gyrus, Fusiform Gyrus, Middle Occipital Gyrus, Lingual Gyrus, Cerebellum, Calcarine, Cerebellum	1803	-34	-90	-12	7.16
R	Occipital	Inferior Occipital Gyrus, Lingual Gyrus, Calcarine, Middle	678	24	-98	-12	6.92

		Occipital Gyrus, Fusiform Gyrus, Superior Occipital Gyrus					
L, R	Occipital	Precuneus, Precuneus, Cuneus, Posterior Cingulum, Cuneus, Outsiede, Posterior Cingulum, Superior Occipital Gyrus, Middle Cingulum, Superior Occipital Gyrus	1083	0	-38	26	4.88
L	Occipital	Angular Gyrus, Middle Occipital Gyrus	119	-44	-66	34	4.48
R		Caudate, Pallidum, Putamen	140	18	-4	6	4.08
T2							
R	Occipital	Inferior Occipital Gyrus, Lingual Gyrus, Calcarine, Middle Occipital Gyrus, Fusiform Gyrus, Superior Occipital Gyrus, Cuneus, Cerebellum, Cerebellum	1242	24	-98	-8	9.42
L	Occipital	Middle Occipital Gyrus, Inferior Occipital Gyrus, Fusiform Gyrus, Lingual Gyrus, Cerebellum, Calcarine, Cerebellum	1975	-26	-96	-8	8.02
L	Occipital, Parietal	Superior Parietal Gyrus, Inferior Parietal Gyrus, Middle Occipital Gyrus, Superior Occipital Gyrus, Angular	582	-26	-68	44	5.47
R	Occipital, Parietal	Superior Parietal Gyrus, Superior Occipital Gyrus, Angular Gyrus, Middle Occipital Gyrus, Inferior Parietal Gyrus	468	28	-68	44	4.72

Supplementary Table S 2.2 Atlas-based mean Intraclass Correlation (ICC) values for the anatomical regions specified in the aal atlas and mean ICC values for the ventral and dorsal striatum regions of interest mask that were built according to the definition of Schacht et al. (2011), in order to allow comparability between studies.

Group	IWT N=21	CET I N=29	CET II N=42	CET+DCS N=32	NTX N=20
Paradigm	ALCUEPV	ALCUE	ALCUEPV	ALCUE	ALCUE
Amygdala_L 4201	0.29	0.19	0.11	-0.15	0.18
Amygdala_R 4202	0.09	-0.03	0.00	-0.21	0.28
Angular_L 6221	0.20	0.27	0.23	0.14	0.08
Angular_R 6222	0.17	0.17	0.15	0.05	0.08
Calcarine_L 5001	0.30	0.25	0.10	0.09	**0.46**
Calcarine_R 5002	0.30	0.28	-0.01	0.05	**0.42**
Caudate_L 7001	0.06	0.13	0.05	0.09	0.13
Caudate_R 7002	0.15	0.16	-0.05	0.02	0.30
Cerebelum_L 9081	0.37	0.35	-0.10	0.02	0.25
Cerebelum_R 9082	0.14	0.02	-0.13	-0.23	0.22
Cerebelum_L 9021	-0.02	-0.11	-0.15	0.13	0.34
Cerebelum_R 9022	-0.10	-0.16	-0.20	0.05	**0.41**
Cerebelum_L 9031	0.17	0.11	-0.16	-0.01	0.25
Cerebelum_R 9032	0.05	0.01	-0.12	-0.01	0.21
Cerebelum_L 9041	0.17	0.17	0.01	-0.05	0.23
Cerebelum R 9042	0.13	0.05	0.01	-0.13	0.29
Cerebelum_L 9051	0.21	-0.11	0.00	0.00	0.19
Cerebelum_R 9052	0.03	0.09	0.05	0.17	0.06
Cerebelum_L 9061	0.19	0.08	-0.06	-0.28	0.18

Cerebelum_R 9062	0.20	0.07	0.22	-0.21	0.16
Cerebelum_L 9071	0.18	0.01	0.20	-0.21	0.20
Cerebelum R 9072	0.13	0.12	0.01	-0.14	0.19
Cerebelum_Crus_L 9001	0.21	0.16	0.12	-0.12	0.26
Cerebelum_Crus_R 9002	0.22	0.17	0.07	-0.10	0.30
Cerebelum_Crus_L 9011	0.33	0.10	-0.04	-0.14	0.18
Cerebelum_Crus_R 9012	0.03	0.13	0.04	-0.13	0.20
Cingulum_Ant_L 4001	0.20	0.09	0.13	-0.12	0.24
Cingulum_Ant_R 4002	0.24	0.08	0.11	-0.14	0.20
Cingulum_Mid_L 4011	0.15	0.11	-0.06	0.11	0.18
Cingulum_Mid_R 4012	0.15	0.09	-0.03	0.05	0.13
Cingulum_Post_L 4021	0.23	0.29	0.21	0.12	0.05
Cingulum_Post_R 4022	0.13	0.31	0.16	0.11	0.02
Cuneus_L 5011	**0.46**	0.33	0.14	0.08	0.19
Cuneus_R 5012	0.35	0.29	0.09	0.00	0.24
Frontal_Inf_Oper_L 2301	0.29	0.18	0.22	0.05	0.27
Frontal_Inf_Oper_R 2302	0.29	0.13	0.14	0.03	0.00
Frontal_Inf_Orb_L 2321	0.25	0.21	0.17	0.15	0.21
Frontal_Inf_Orb_R 2322	0.28	0.18	0.07	0.05	0.12
Frontal_Inf_Tri_L 2311	0.16	0.28	0.21	-0.06	0.28
Frontal_Inf_Tri_R 2312	0.26	0.09	0.13	-0.05	0.01
Frontal_Med_Orb_L 2611	0.05	0.10	0.24	-0.08	0.20
Frontal_Med_Orb_R 2612	0.10	0.15	0.14	-0.02	0.10
Frontal_Mid_L 2201	0.28	0.21	-0.01	-0.05	0.31
Frontal_Mid_Orb_L 2211	0.23	0.23	0.28	0.12	0.28
Frontal_Mid_Orb_R 2212	0.24	0.25	0.11	-0.03	0.14
Frontal_Mid_R 2202	0.25	0.19	0.03	-0.10	0.09

Frontal_Sup_L 2101	0.22	0.10	0.02	-0.04	0.28
Frontal_Sup_Medial_L 2601	0.25	0.06	0.02	-0.08	0.20
Frontal_Sup_Medial_R 2602	0.16	0.12	0.03	-0.10	0.16
Frontal_Sup_Orb_L 2111	0.20	0.09	0.30	-0.12	0.05
Frontal_Sup_Orb_R 2112	0.19	0.24	0.03	-0.10	0.10
Frontal_Sup_R 2102	0.14	0.15	0.01	-0.09	0.18
Fusiform_L 5401	0.21	0.30	0.25	0.21	0.30
Fusiform_R 5402	0.19	0.20	0.09	0.24	0.27
Heschl_L 8101	**0.57**	0.19	0.12	-0.08	0.13
Heschl_R 8102	**0.42**	0.15	-0.01	-0.12	0.22
Hippocampus_L 4101	0.24	0.30	0.08	0.00	0.21
Hippocampus_R 4102	0.27	0.06	0.01	-0.01	0.17
Insula_L 3001	0.31	0.14	0.06	0.00	0.11
Insula_R 3002	0.21	0.12	0.03	-0.10	0.11
Lingual_L 5021	0.26	0.34	0.14	0.11	0.30
Lingual_R 5022	0.37	0.30	0.04	0.00	0.30
Occipital_Inf_L 5301	0.33	0.33	0.36	0.35	0.33
Occipital_Inf_R 5302	0.32	0.37	0.28	0.29	0.32
Occipital_Mid_L 5201	0.30	0.32	0.27	0.22	0.38
Occipital_Mid_R 5202	0.25	0.32	0.21	0.21	0.32
Occipital_Sup_L 5101	**0.40**	**0.43**	0.16	0.06	0.25
Occipital_Sup_R 5102	0.31	0.36	0.10	0.05	0.22
Olfactory_L 2501	0.25	0.24	0.16	0.00	0.11
Olfactory_R 2502	0.16	0.22	-0.06	-0.03	0.09
Pallidum_L 7021	0.17	-0.04	0.16	-0.02	**0.41**
Pallidum_R 7022	0.32	0.13	-0.03	0.01	0.38
Paracentral_Lobule_L 6401	0.38	0.22	-0.07	-0.05	0.12

Paracentral_Lobule_R 6402	0.30	0.21	-0.13	0.02	0.20
Para_Hippocampal_L 4111	0.18	0.08	0.01	0.03	0.15
Para_Hippocampal_R 4112	0.14	-0.06	-0.09	0.03	0.11
Parietal_Inf_L 6201	0.17	0.33	0.01	0.19	0.30
Parietal_Inf_R 6202	0.10	0.38	-0.01	0.12	0.19
Parietal_Sup_L 6101	0.39	0.30	0.03	0.09	0.38
Parietal_Sup_R 6102	0.28	0.24	0.06	0.03	0.29
Postcentral_L 6001	0.28	0.22	0.00	0.00	0.12
Postcentral_R 6002	0.30	0.10	-0.03	0.03	0.09
Precentral_L 2001	0.16	0.22	0.11	-0.02	0.23
Precentral_R 2002	0.22	0.19	0.00	-0.10	0.06
Precuneus_L 6301	0.24	0.32	0.11	0.11	0.07
Precuneus_R 6302	0.07	0.30	0.08	0.12	0.10
Putamen_L 7011	0.31	0.04	0.11	-0.01	0.33
Putamen_R 7012	0.23	0.07	0.10	-0.14	0.35
Rectus_L 2701	0.15	0.22	0.25	-0.03	0.10
Rectus_R 2702	0.19	0.17	0.12	-0.09	0.17
Rolandic_Oper_L 2331	**0.51**	0.13	0.13	-0.11	0.11
Rolandic_Oper_R 2332	0.39	0.19	0.04	-0.24	0.16
Supplementary_Motor_Area_L 2401	0.24	0.16	-0.04	-0.07	0.03
Supplementary_Motor_Area_R 2402	0.22	0.22	-0.05	-0.09	0.03
SupraMarginal_L 6211	0.27	0.18	-0.09	-0.02	0.28
SupraMarginal_R 6212	0.27	0.04	-0.04	0.02	0.00
Temporal_Inf_L 8301	0.25	0.18	0.21	0.01	0.21
Temporal_Inf_R 8302	0.15	0.22	0.08	0.14	0.23

Temporal_Mid_L 8201	0.28	0.15	0.20	0.06	0.26
Temporal_Mid_R 8202	0.16	0.11	0.14	0.05	0.22
Temporal_Pole_Mid_L 8211	0.36	-0.08	-0.09	-0.03	0.04
Temporal_Pole_Mid_R 8212	0.25	0.19	0.06	-0.05	0.02
Temporal_Pole_Sup_L 8121	0.11	0.11	0.02	-0.01	0.17
Temporal_Pole_Sup_R 8122	0.02	0.16	-0.03	-0.09	0.07
Temporal_Sup_L 8111	0.37	0.18	0.10	-0.03	0.12
Temporal_Sup_R 8112	0.27	0.16	0.02	-0.05	0.14
Thalamus_L 7101	0.32	0.23	0.09	-0.10	0.19
Thalamus_R 7102	0.35	0.25	0.06	-0.17	0.21
Vermis_9100	0.17	0.00	-0.19	-0.01	0.30
Vermis_9110	0.02	0.05	-0.08	0.16	0.26
Vermis 9120	0.07	0.11	-0.01	-0.12	0.21
Vermis_9130	0.03	0.10	0.12	-0.15	0.20
Vermis_9140	-0.14	0.08	0.14	-0.31	0.36
Vermis_9150	-0.18	0.09	0.00	-0.29	0.22
Vermis 9160	0.01	0.12	0.00	-0.24	0.18
Vermis_9170	0.11	-0.16	0.00	-0.07	0.06
DS_L (Schacht et al. 2011)	0.19	-0.04	0.01	-0.10	-0.04
DS_R (Schacht et al. 2011)	0.30	-0.01	-0.11	-0.08	0.06
VS_L (Schacht et al. 2011)	0.09	0.33	0.06	0.33	0.10
VS_R (Schacht et al. 2011)	0.25	0.36	-0.21	0.16	0.29

VS=Ventral Striatum, DS=Dorsal Striatum, L=left hemisphere, R=right hemisphere, Ant=Anterior, Post=Posterior, Mid=Middle, Sup=Superior, Inf=Inferior, numbers indicate the number of the aal atlas, ALCUE = version 1 of the alcohol cue-reactivity task, ALCUEPV = version 2 of the alcohol cue-reactivity task without a rating phase

Supplementary Table S 2.3 Jaccard and Dice coefficients for the three task condition contrasts alcohol-neutral, alcohol and neutral, illustrating the proportion of overlapping significant voxels between the and second fMRI session (for the thresholds p<0.001 and p<0.01, defining super-threshold activation).

Jaccard Coefficient							
A] Threshold p < 0.01							
Difference Contrast: alcohol vs. neutral							
Group		IWT	CET I	CET II	CET + DCS	IWT + NTX	Total
	M	0,07	0,03	0,10	0,06	0,09	**0,07***
	SD	0,11	0,06	0,11	0,09	0,09	0,10
Contrast: alcohol							
Group		IWT	CET I	CET II	CET + DCS	IWT + NTX	Total
	M	0,51	0,50	0,52	0,58	0,52	**0,53***
	SD	0,20	0,18	0,17	0,13	0,17	0,17
Contrast: neutral							
Group		IWT	CET I	CET II	CET + DCS	IWT + NTX	Total
	M	0,51	0,50	0,55	0,58	0,52	**0,54***
	SD	0,21	0,18	0,17	0,13	0,18	0,17
B] Threshold p < 0.001							
Difference Contrast: alcohol vs. neutral							
Group		IWT	CET I	CET II	CET + DCS	IWT + NTX	Total
	M	0,09	0,05	0,11	0,08	0,10	**0,08***
	SD	0,12	0,07	0,10	0,10	0,09	0,10
Contrast: alcohol							
Group		IWT	CET I	CET II	CET + DCS	IWT + NTX	Total
	M	0,46	0,46	0,48	0,53	0,49	**0,48***
	SD	0,18	0,14	0,15	0,10	0,14	0,14
Contrast: neutral							

Group		IWT	CET I	CET II	CET + DCS	IWT + NTX	Total
	M	0,46	0,44	0,49	0,52	0,48	**0,48***
	SD	0,17	0,16	0,15	0,14	0,14	0,15
Dice Coefficient							
C] Threshold p < 0.01							
Difference Contrast: alcohol vs. neutral							
Group		IWT	CET I	CET II	CET + DCS	IWT + NTX	Total
	M	0,12	0,06	0,17	0,09	0,15	**0,12***
	SD	0,17	0,11	0,17	0,14	0,15	0,15
Contrast: alcohol							
Group		IWT	CET I	CET II	CET + DCS	IWT + NTX	Total
	M	0,65	0,65	0,67	0,72	0,67	**0,67***
	SD	0,23	0,19	0,19	0,11	0,19	0,18
Contrast: neutral							
Group		IWT	CET I	CET II	CET + DCS	IWT + NTX	Total
	M	0,64	0,65	0,69	0,72	0,66	**0,68***
	SD	0,24	0,19	0,18	0,11	0,19	0,18
D] Threshold p < 0.001							
Difference Contrast: alcohol vs. neutral							
Group		IWT	CET I	CET II	CET + DCS	IWT + NTX	Total
	M	0,15	0,08	0,18	0,13	0,17	**0,14***
	SD	0,17	0,11	0,15	0,15	0,14	0,15
Contrast: alcohol							
Group		IWT	CET I	CET II	CET + DCS	IWT + NTX	Total
	M	0,61	0,62	0,63	0,68	0,64	**0,64***
	SD	0,21	0,15	0,17	0,09	0,14	0,16
Contrast: neutral							
Group		IWT	CET I	CET II	CET + DCS	IWT + NTX	Total

	M	0,61	0,60	0,64	0,67	0,63	**0,63***
	SD	0,19	0,18	0,17	0,12	0,15	0,16

* = significant main effect of task contrast condition (alcohol-neutral, alcohol, neutral) at $p < 0.05$, such that values for the difference contrasts (alcohol-neutral) significantly differed from the other two conditions

Supplementary Table S 2.4 Results of the data simulation on the magnitude of the correlations between the constituting contrast conditions (V_1 and V_2) and external variables (correlating with the constituting contrast conditions to V_3=0.1, V_4=0.2, V_5=0.3, V_6=0.4, V_7=0.5, V_8=0.6, V_9=0.7, V_{10}=0.8) and between the difference score (V_1-V_2) and the external variables. The absolute difference increases for higher correlations between the constituting task conditions and the external variables, while the percent difference remained stable at about 33 to 34% across all correlation magnitudes.

	V_1	V_2	V_1-V_2	**Absolute Difference** (V_1,V_2 vs. V_1-V_2)	**Deviation in %**
V_3	0.100	0.099	0.066	0.034	33.668
V_4	0.199	0.199	0.132	0.067	33.668
V_5	0.298	0.298	0.197	0.101	33.893
V_6	0.398	0.398	0.264	0.134	33.668
V_7	0.499	0.498	0.329	0.170	34.002
V_8	0.598	0.598	0.396	0.202	33.779
V_9	0.699	0.699	0.462	0.237	33.906
V_{10}	0.799	0.798	0.527	0.272	34.001

Supplementary Figures

Supplementary Figure S 2.1 Depiction of brain areas that show moderate to good reliability or the alcohol condition contrast (Intraclass correlation [ICC] > 0.4) for the different study groups (A to E, left column) and histograms depicting the distribution of ICC values and the mean and median for the different study groups (right column).

Brain regions with ICC values > 0.4 (“moderate“)

Histograms of ICC values

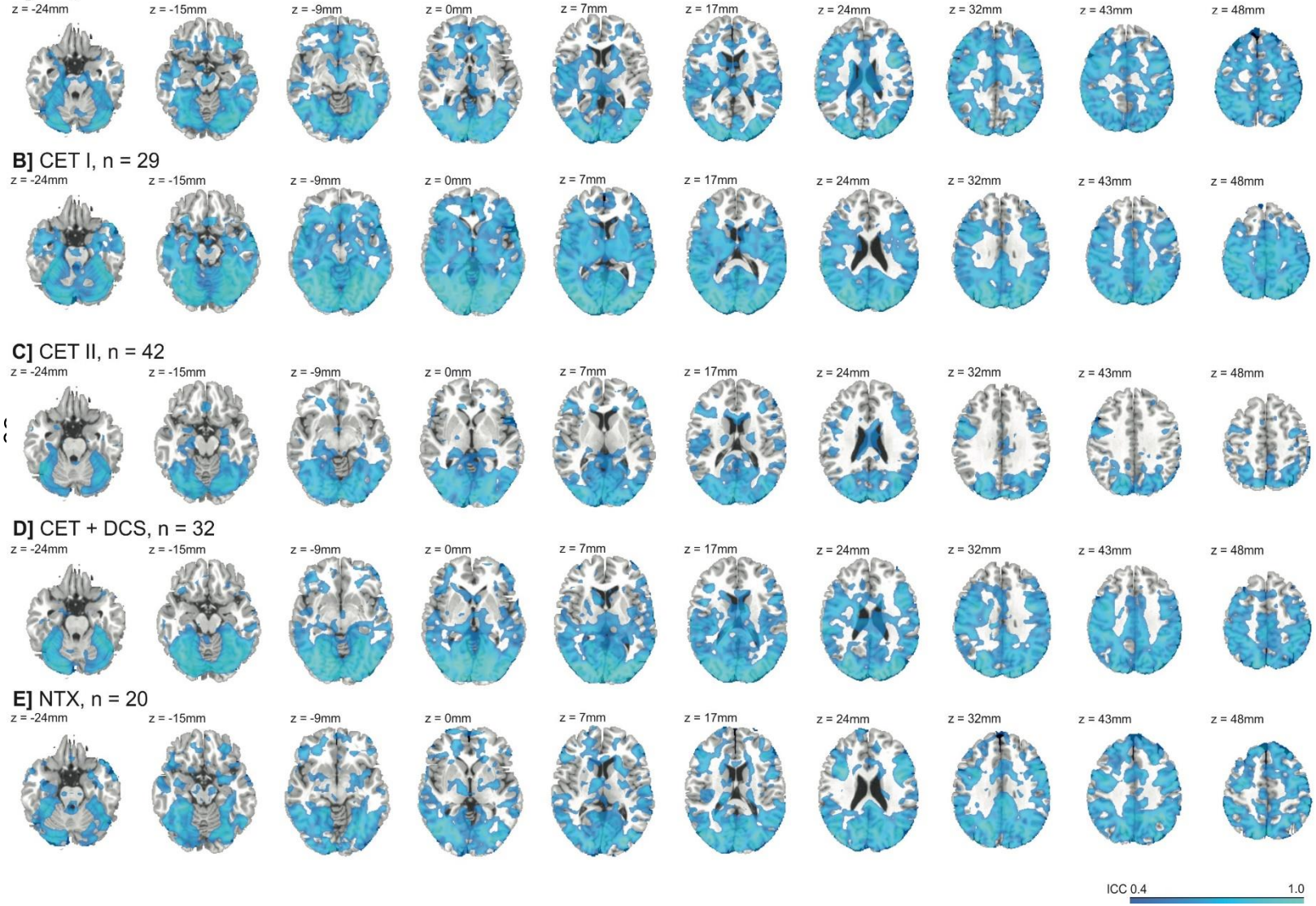
A] IWT, n = 21

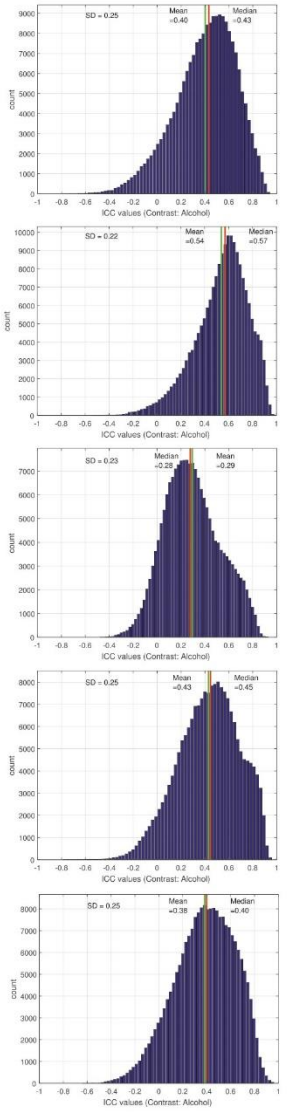
B] CET I, n = 29

C] CET II, n = 42

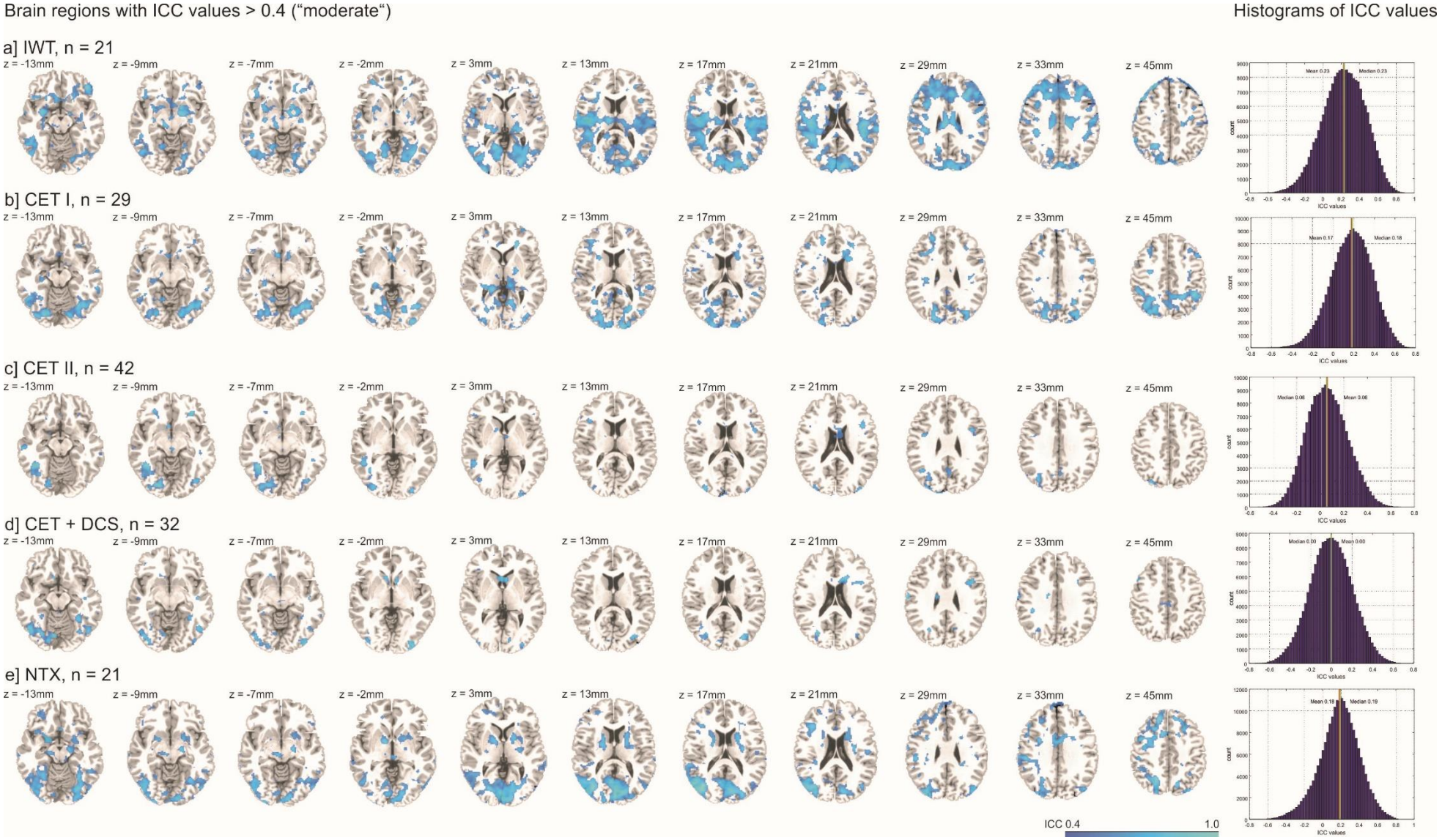
D] CET + DCS, n = 32

E] NTX, n = 20

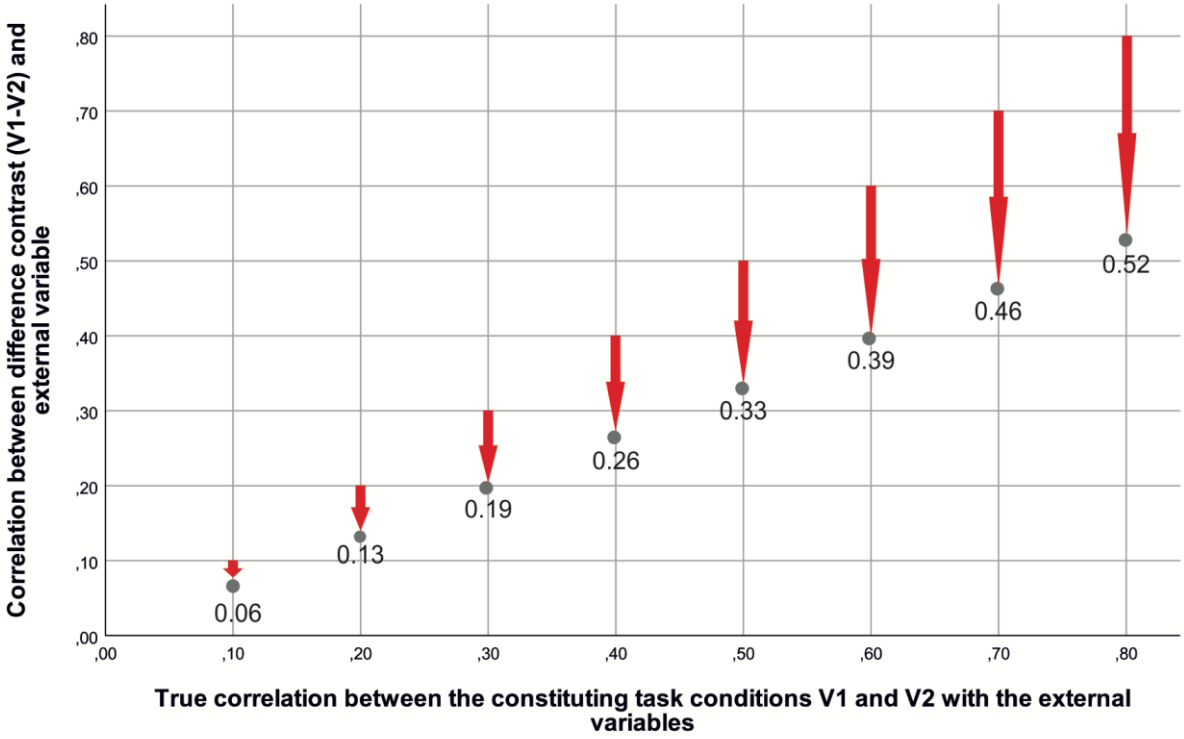




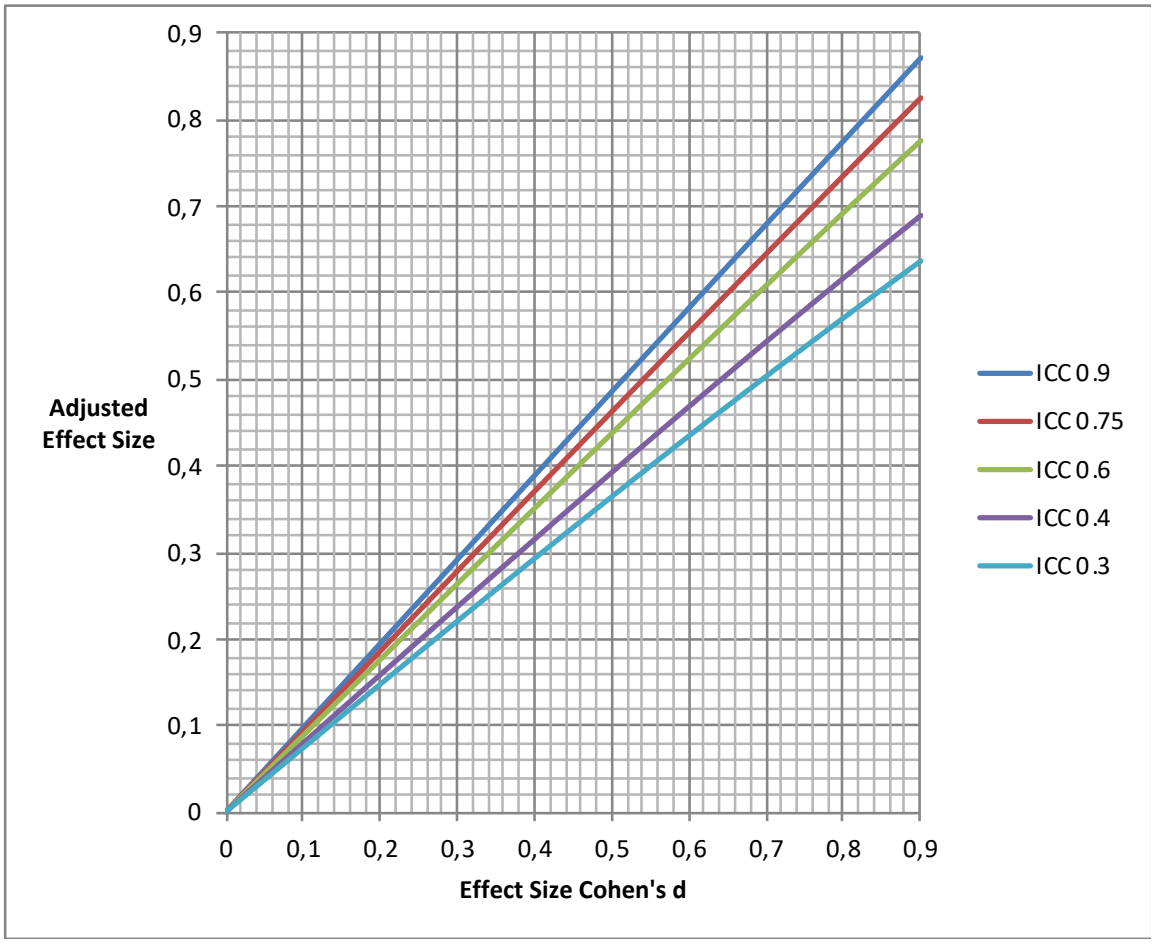
Supplementary Figure S 2.2 Depiction of brain areas that show moderate to good reliability for the difference contrast alcohol-neutral (Intraclass correlation [ICC] > 0.4) for the different study groups (A to E, left column) and histograms depicting the distribution of ICC values and the mean and median for the different study groups (right column).


Brain regions with ICC values > 0.4 (“moderate“)
Histograms of ICC values
a] IWT, n = 21
b] CET I, n = 29
c] CET II, n = 42
d] CET + DCS, n = 32
e] NTX, n = 21
z = -13mm
z = -9mm
z = -7mm
z = -2mm
z = 3mm
z = 13mm
z = 17mm
z = 21mm
z = 29mm
z = 33mm
z = 45mm
Mean 0.23
Median 0.23
Mean 0.17
Median 0.18
Median 0.06
Mean 0.06
Median 0.00
Mean 0.00
Mean 0.18
Median 0.19
count
ICC values
ICC 0.4
1.0

Supplementary Figure S 2.3 Depiction of the results of the data simulation illustrating the lower correlation between the simulated external variables (modeled after the OCDS scores, correlating with the constituting contrast conditions V_1 and V_2 to 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8) and the difference score (V_1-V_2).



Supplementary Figure S 2.4 Illustration of the dependence between effect size estimates (Cohen's d), reliability and adjusted Cohen's d approximating the "true" population effect. Values show that the effect sizes are substantially over-estimated when studies rely on un-reliable measures.



2.2 Study 2 - Reliability of neural food-cue reactivity in participants with obesity undergoing bariatric surgery: a 26-week longitudinal fMRI study[2]

2.2.1 Abstract

Obesity is highly prevalent worldwide and results in a high disease burden. The efforts to monitor and predict treatment outcome in participants with obesity using functional magnetic resonance imaging (fMRI) depends on the reliability of the investigated task-fMRI brain activation. To date, no study has investigated whole-brain reliability of neural food cue-reactivity. To close this gap, we analyzed the longitudinal reliability of an established food cue-reactivity task.

Longitudinal reliability of neural food-cue induced brain activation and subjective food craving ratings over three fMRI sessions (T0: two weeks before surgery, T1: eight weeks and T2: 24 weeks after surgery) were investigated in N=11 participants with obesity. We computed an array of established reliability estimates, including the intraclass correlation (ICC), the Dice and Jaccard coefficients and similarity of brain activation maps.

The data indicated good reliability (ICC > 0.6) of subjective food craving ratings over 26 weeks and excellent reliability (ICC > 0.75) of brain activation signals for the contrast of interest (food>neutral) in the caudate, putamen, thalamus, middle cingulum, inferior, middle and superior occipital gyri, and middle and superior temporal gyri and cunei. Using similarity estimates, it was possible to re-identify individuals based on their neural activation maps (73%) with a fading degree of accuracy, when comparing fMRI sessions further apart.

[2] Bach P, Grosshans M, Koopmann A, Kienle P, Vassilev G, Otto M, Bumb JM, Kiefer F. Reliability of neural food cue-reactivity in participants with obesity undergoing bariatric surgery: a 26-week longitudinal fMRI study. Eur Arch Psychiatry Clin Neurosci. 2021 Aug;271(5):951-962. doi: 10.1007/s00406-020-01218-8. Epub 2020 Dec 17. PMID: 33331960; PMCID: PMC8236041.

The results show excellent reliability of task-fMRI neural brain activation in several brain regions. Current data suggests that fMRI-based measures might indeed be suitable to monitor and predict treatment outcome in participants with obesity undergoing bariatric surgery.

2.2.2 Introduction

Obesity affects more than 650 million people worldwide (WHO, 2018). Overweight has been identified as a major cause of cardiovascular diseases, diabetes, musculoskeletal disorders as well as several types of cancer (Bhaskaran et al., 2014). The assessment of behavioral and neural responses towards food cues has received some interest in the last decade as a tool to investigate the neurobiological basis of obesity (Harding et al., 2018; Kerem et al., 2019). A recent meta-analysis on food cue-reactivity concluded that across 45 published reports the overall effect of food cue-reactivity and craving on outcomes in patients was of medium size (r=0.3) with a large variability across studies. Authors concluded that food cue exposure and the experience of craving have a significant influence on and contribute to eating behavior and weight gain (Boswell & Kober, 2016). Functional magnetic resonance imaging (fMRI) was used to identify the neural correlates of food craving, food perception, and food intake. Structures implicated in food-intake regulation include the anterior insula, inferior frontal and orbitofrontal cortices, the medial temporal cortex with the amygdala and parahippocampus, as well as the nucleus accumbens, and visual cortices (Benarroch, 2010). In the last years, efforts were undertaken to establish neural predictors for eating behavior and treatment response. Some studies reported significant associations between neural responses to food cues and weight loss during treatment, but overall the studies report heterogeneous findings (Huerta et al., 2014). The inconsistencies in study results demand for an investigation of the reliability and robustness of the applied food cue-reactivity task, because the possibility to establishing meaningful and robust associations between neural brain responses during food cue presentation and any behavioral or clinical variable critically depends on the reliability of the investi-

gated task-fMRI brain activation. Previous studies have demonstrated substantial variability in findings of food cue-reactivity studies. Although brain responses to visual food cues in participants with obesity have been found to have relatively good mean-level reproducibility, they had poor within-subject test-retest reliability (R Drew Sayer et al., 2016). Several factors were associated with the heterogeneity in findings, including different expression of the fat mass and obesity-associated genes (e.g. FTO) (Karra et al., 2013; Rapuano et al., 2017; Wiemerslage et al., 2016), fasted state vs. glucose ingestion prior to fMRI (Heni et al., 2014) and divergent characteristics of the individual study designs, including the structure, timing and stimuli used during the food cue-reactivity fMRI task. Further, there are clear individual differences in food preferences that were associated with additional variance across studies (Van Der Laan & Smeets, 2015). Additionally, small sample sizes and a lack of power were related to inconsistencies between studies (Button et al., 2013). Moreover, a study comparing the results of 70 different teams analyzing the same dataset, revealed significant variability in the analysis of the same fMRI food cue-reactivity dataset depending on the researcher's decision to use a certain the statistical software (e.g. SPM vs. FSL vs. AFNI) or statistical method (parametric vs. non-parametric) as well as the applied smoothing kernel (Botvinik-Nezer et al., 2020). The results highlight the need for better standardization of the food stimuli and fMRI task designs and the additional data that is collected on participant's state (hunger, mood, hormones etc.) and personal characteristics that may be used to control for confounding effects in the analyses. The aforementioned findings emphasize the importance of establishing standardized food cue-reactivity paradigms, study protocols and analysis workflows. To this end, guidelines for good practice in food cue-reactivity neuroimaging studies were proposed. According to these guidelines, researchers planning fMRI studies should take special care to: power calculation, hunger state and related factors, personal characteristics, the selection of food-related stimuli, setting well-considered statistical thresholds for whole brain analyses, minimizing the risk of movement artifacts, analysis of prospective designs as well as predictive modelling. Moreover, the authors suggest to pre-register planned studies and to share the data obtained (Smeets et al.,

2019). In doing so, it would be possible to ensure reproducibility of results across cue-reactivity studies (Smeets et al., 2019).

To date, there is no study that investigated whole-brain reliability of food cue-induced brain activation. To our knowledge, only a single fMRI study investigated the longitudinal reliability of extracted mean brain activation during food cue processing over a mean period of 18 days (3-35 days), which is short considering follow-up periods of clinical studies that run over months. Additionally, reliability was only assessed in a selected range of a priori defined regions of interest (bilateral insula, amygdala, orbitofrontal cortex, caudate and putamen) (R. Drew Sayer et al., 2016). The authors reported that in their dataset, only the left orbitofrontal cortex response showed fair reliability, while all other regions of interest showed poor reliability. The authors also stated that the large inter-individual range of days between the two assessment sessions might have limited reliability in their study. Additionally, previous research highlighted that low reliability in fMRI studies might also be associated to the computation of difference scores or difference contrasts, where one condition is subtracted from the other. For example, regarding the food cue-reactivity tasks, it is common to subtract the brain activation during food picture blocks from activation during neutral picture blocks. However, in the case of a high correlation between the constituting conditions of a difference score, the resulting reliability of that score is limited, because much of the shared "true" variance is removed, while the measurement errors are added (Infantolino et al., 2018; Peter et al., 1993). To date however, no study investigated whole brain reliability of food-cue induced brain responses over a longer period of time and, importantly, no study to date investigated reliability in samples of patients undergoing surgery. This, however, seems relevant to the ongoing efforts to establish predictors and biomarkers for treatment efficacy in obesity. In this context, it is necessary to determine the reliability of food cue-reactivity in clinical populations undergoing treatment, because only this way the robustness and suitability of cue-reactivity as a biomarker in obesity can be assessed. Hence, we conducted our analyses in a clinical population undergoing surgery, as this sample reflects a sample for whom biomarkers should be established to predict and monitor treatment outcomes

using fMRI biomarkers. Hence, we set out to assess the reliability of neural food-cue reactivity in a longitudinal dataset of individuals with obesity over three neuroimaging assessments that were scheduled two weeks before bariatric surgery, and eight and twenty-four weeks after surgical intervention. We used an unrestricted whole-brain approach and a set of complementary measures for fMRI reliability, aimed at determining the global and local reliability of the difference contrast (food-neutral) and of the constituting food and neutral picture conditions. Additionally, we compared the reliability of food cue-reactivity to the reliability of commonly applied subjective craving measures that were measured during the fMRI session.

2.2.3 Methods

Participants

Current analyses were conducted on a dataset of N=11 individuals with obesity of whom fMRI task data was available for three time points and that were part of a larger longitudinal clinical study, including a total of N=26 participants with obesity, of whom however only the N=11 participants met the inclusions criteria for undergoing fMRI scanning (e.g. absence of metal implants, claustrophobia and waist circumference < 160 cm (due to the scanner diameter). The clinical data of the of the whole study group are reported elsewhere. In short, patients showed a percent total weight loss after surgery (%TWL) from T0 to T2 of 23.8 %TWL after Roux-en-Y gastric bypass (n=21) and 12.7 %TWL after sleeve gastrectomy (n=5) with no significant difference between both procedures ($p = 0.126$). There were also significant reductions of resting heart rate, fasting plasma glucose levels and depressive symptoms (all $p < 0.001$). Only individuals with obesity that already decided to receive bariatric surgery were recruited for this study. The study procedure was approved by the local ethics committee and all participants provided written informed consent.

Individuals with obesity undergoing fMRI had to meet the following inclusion criteria: i) age between 18 and 65 years, ii) BMI (kg/m^2) > 35 (i.e. ≥ grade 2 obesity), iii) a waist

circumference < 160 cm (limited by scanner diameter), iv) the capacity to give informed consent, v) no history or current diagnosis of any psychiatric, neurological, neoplastic or untreated endocrine illnesses (with the exception of nicotine addiction), and no current intake of any centrally acting psychoactive or anti-obesity medications (i.e. sedatives, antipsychotics, including long-acting injectable antipsychotics, antidepressants, opioid analgesics as well as DPP (dipeptidyl peptidose IV) inhibitors and GLP (Glucagon-like peptide)-1 antagonists, vi) all participants with a history of surgical interventions in the gastrointestinal system or contraindications to fMRI scanning (e.g. metal implants), and pregnant or breast-feeding females were excluded.

Twenty-six individuals (17 females and 9 males, mean age 41 ± 12 years, mean BMI 46 ± 6 kg/m^2) were eligible for analyses (demographics, bariatric surgery, blood analyses as well as behavioral data) and included in the study. Of these 26 individuals, 21 received Roux-en-Y gastric bypass and 5 sleeve gastrectomy. Imaging data could be obtained for 11 obese individuals (10 individuals received Roux-en-Y gastric bypass and 1 sleeve gastrectomy; 15 individuals had to be excluded due to the fact that they did not fit the scanner.

Procedures

To (Two weeks before bariatric surgery)

During the first assessment session, sociodemographic data, information on internal and neurological disorders, as well as information on eating habits was collected. In addition, participants were screened for any psychiatric comorbidities using the Structured Clinical Interviews for DSM-IV, SKID-I, (H.U. Wittchen et al., 1997). Additionally, a urine drug screening, and in females a pregnancy test was conducted.

fMRI scanning was performed between noon and 3 PM. All participants received a standardized breakfast of 500 kcal (2093 kJ) 6 hours before fMRI scanning and did not eat until the scanning. Subsequently, participants completed a series of questionnaires including the Beck Depression Inventory (BDI, (Beck et al., 1961), the Fagerstrom Test for Cigarette Dependence (FTCD (Fagerstrom, 2012) as well as the Yale Food Addiction Scale (YFAS) (Gearhardt et al., 2009).

T1 and T2 (Eight and 24 weeks after bariatric surgery)

At both time points, participants were examined medically, urine drug screenings, and in females a pregnancy test were performed. Moreover, possible changes in medication were documented. MRI measurements were performed at both time points using the same procedures and tasks as during the first scanning session.

Imaging procedure

fMRI food cue-reactivity task

All patients included in the current analyses underwent three different imaging sessions. During these sessions, patients laid in the scanner wearing MRI-compatible goggles, on which sets of visual food and neutral stimuli were presented using a block design. The task consisted of a total of 18 blocks of food stimuli and 12 blocks of neutral stimuli. Each block comprised of a series of five food or neutral pictures. Food stimuli were further divided in three categories: salty high-calorie, sweet high-calorie, low-calorie, yielding 6 blocks for each category. All stimuli were shown for 4 seconds (i.e. 20 seconds per block) in a pseudo-randomized order. Participants were instructed to closely watch each picture and were informed that they will be asked to rate their subjective craving. In-between each picture block, patients were asked to rate their current craving for food on a visual analogue scale (VAS) that ranged from 0 - "very weak" to 100 - "very strong". The fMRI took 18 minutes. Food stimuli chosen were rated according to their ability to induce food craving by 44 voluntary participants at our institution (Grosshans et al., 2012) and neutral cues were taken from the International Affective Picture Series (Lang et al., 1999).

fMRI acquisition and pre-processing

A total of 453 images T2*-weighted, echo planar images covering the entire brain were acquired during the food cue task using a 3-T whole-body tomography scanner (MAGNETOM Trio with TIM technology; Siemens). Imaging parameters were: repetition time = 2.41 seconds, echo time = 25 milliseconds, flip angle = 80°, number of slices = 42, slice thickness = 2 mm, voxel-gap = 1 mm, voxel dimensions = 3 × 3 × 3 mm^3, field

of view = 192 × 192 mm2, in-plane resolution = 64 × 64. The short echo time and the 30° flip angle to anterior commissure–posterior commissure orientation was chosen to minimize susceptibility artefacts. Stimuli were presented using Presentation software (version 9.9, Neurobehavioral Systems Inc.) and MRI-compatible goggles (MRI Audio/Video Systems; Resonance Technology Inc., CA).

Functional-imaging data was processed and analyzed using SPM8 and SPM12. The first 5 scans were excluded from imaging analyses to avoid any artefacts caused by the effects of magnetic saturation. All images were realigned spatially (movement was considered excessive with > 2mm translation or >2° rotation), normalized to a standardized EPI template from MNI (Montreal Neurological Institute, Quebec, Canada), and smoothed using an isotropic Gaussian kernel for group analyses (full width at half maximum: 8 mm).

Food cue-reactivity imaging data was analyzed by modelling the different task conditions (food with the subcategories salty high-calorie, sweet high-calorie, low-calorie and neutral) as explanatory variables within a general linear model in SPM implementing the movement parameters as nuisance variables. Individual contrast images (food cues > neutral cues) were computed for each individual and then included into following second-level analyses in SPM. Nicotine consumption (categorical) was considered as covariate, because previous work indicated that nicotine modulates food–cue reactivity (Kroemer et al., 2013). In order to satisfy a family-wise error rate correction of pFWE < .05, we determined a combined height ($p < .001$) and extent ($k \geq 103$) threshold by running 10.000 Monte Carlo simulations using AlphaSim as implemented in the Neuroelf analysis package (www.neuroelf.net) (Bennett et al., 2009), (estimated smoothness was x/y/z = 10.13/9.86/10.33 mm) (Eklund et al., 2016).

Reliability analyses

We investigated the reliability of subjective food craving ratings (i.e. mean craving for food – mean craving for neutral stimuli during the 1st, 2nd and 3rd assessment session, in order to correspond to the fMRI task contrast "food – neutral") over the three fMRI sessions by computing the intraclass correlation coefficients using a two-way, mixed effects model in IBM SPSS (version 25.0). Additionally, whole-brain longitudinal reliability of individual brain responses to food stimuli over the three imaging sessions by computing measures of local and global reliability using the fmreli toolbox for SPM12 by Kroemer, Frohner and colleagues (Frohner et al., 2019) (https://github.com/nkroemer/reliability). Analyses were conducted on the whole brain without a-priori restrictions to specific regions of interest.

Jaccard and Dice coefficients

We computed the modified Jaccard coefficient, a common measure in fMRI reliability studies between the three different time points for the difference contrast food > neutral and the constituting contrasts (i.e. food and neutral separately). It is defined as the size of the intersection divided by the size of the union of the voxel sets and computed as follows:

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

The Jaccard coefficient can be interpreted as the percentage of overlapping significant voxels above a predefined statistical threshold (e.g. $p < 0.001$) within all significant voxels (Jaccard, 1902; Maitra, 2010).

Additionally, we computed the Dice coefficient for the three different contrasts and scanning time points. It is calculated as the number of super-threshold voxels that overlap between sessions divided by the average number of significant voxels across sessions:

$$Dice(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

The Dice coefficient was introduced to assess the overlap of significant fMRI clusters between scans. It has become an established measure of fMRI data reliability (Rombouts et al., 1997). Both coefficients have values from 0 ("no overlap") to 1 ("perfect overlap") between significant super-threshold voxels. Both measures are, however, limited by the missing consensus on specific values or cut-offs that would differentiate between "poor" and "good" values (Bennett & Miller, 2010). Additionally, the magnitude of both coefficients depends on the statistical threshold used to define what is "active". Studies showed that the reliability of the cluster overlap method decreases, when the significance threshold is increased (Duncan et al., 2009; Rombouts et al., 1998). In the current analyses, we therefore applied a commonly used threshold of $p < 0.001$. Resulting values were imported into the IBM SPSS Statistics software (version 25.0) for further analyses using a repeated measure analysis of variance (ANOVA) model with the factors time (1^{st}, 2^{nd}, 3^{rd}) assessment and task contrast (food, neutral, food > neutral).

Similarity

Secondly, we calculated the within- and between-subject similarity of the fMRI activation maps using the fmreli toolbox (Frohner et al., 2019). Similarity in this context is defined as the resemblance of two activation patterns based on the alignment of high versus low brain activation values across the brain between- and within-subjects (for details see Frohner et al., 2019). The resulting coefficients are correlation coefficients that range from 'perfect' inverse relationship (-1.00) to a 'perfect' direct relationship (1.00). It was suggested that individuals can be successfully identified by their neural activation patterns, if the within-subject similarity exceeds all between-subject association coefficients of the same participant (Finn et al., 2015; Frohner et al., 2019). An advantage of this procedure is that it does not require an a-priori (and potentially arbitrary) statistical threshold.

Intraclass Correlation (ICC)

Thirdly, we estimated voxel-wise reliability of brain activation patterns by computing the intraclass correlation (ICC) coefficients between all three fMRI sessions. The ICC is

used to assess whether the magnitude of activation in each voxel of the brain is stable from test scan to retest scan. Previous work suggested that this measure might be more stringent than other fMRI reliability measures, as it also requires near zero values to be stable over time (Bennett & Miller, 2010). It was suggested that the ICC(3,1) variant is most appropriate for assessing longitudinal fMRI datasets (Ombao et al., 2016). Mathematically, this coefficient sets within-subject variance (σ^2_{within}) in relation to between-subject variance ($\sigma^2_{between}$). We used the ICC(3,1)-type to assess voxel-wise reliability (Shrout & Fleiss, 1979), defined as:

$$ICC = \frac{(\sigma^2{}_{\text{between}} - \sigma^2{}_{\text{within}})}{(\sigma^2{}_{\text{between}} + \sigma^2{}_{\text{within}})}$$

According to Fleiss (1986), ICC coefficients lower than 0.4 represent poor reliability, ICCs between 0.4 and 0.75 represent fair (< 0.6) to good (>0.6) reliability, and ICCs higher than 0.75 represent good to excellent reliability (Fleiss, 1986). We calculated ICC coefficients for every brain voxel, in order to allow identification of brain regions that show high reliability without restriction to predefined regions of interest. However, we were aware that much of the (un-thresholded) brain activation might be unrelated to food cue task and hence would not replicate in its magnitude, resulting in a low overall ICC value. Therefore, we generated thresholded ICC brain maps, to identify brain areas that show good to excellent (ICC > 0.75) reliability and we computed additional atlas-based mean ICC values for a standard set of anatomical brain regions (see below).

Spearman's correlation

In order to assess whether reliability of the common difference contrast food > neutral might be limited by a high correlation between the constituting conditions, we computed the voxel-wise Spearman's correlation coefficients between the three food image categories (i.e. sweet, high caloric, low caloric) and the neutral condition using the fmreli toolbox.

Computation of Atlas-based summary measures

In accordance to previous work (Frohner et al., 2019), we computed the mean ICC for N = 120 anatomical regions specified in the Automatic Anatomic Labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002). The additional atlas-based summary intended to facilitate the assessment of local differences in reliability and identify reliable anatomical ROIs for future analyses. ICC values were extracted using the ROI data extraction routine of the MarsBar software package (http://marsbar.sourceforge.net/) and was imported into SPSS (IBM SPSS Statistics version 25.0) for further analyses.

Group-level fRMI task activation

On a group level, imaging data for every single time point (e.g. 1st, 2nd, 3rd assessment) were analyzed using a one sample t-test, in order to assess the robustness of task main effects (i.e. between condition effects) on group-level brain activation and to determine brain areas that show higher brain activation in response to food cues, compared to neutral cues (contrast: food - neutral). Additionally, we performed analyses of changes in food cue-induced brain responses over time, by setting up a flexible factorial model with the within subject factor time (i.e. 1st, 2nd, 3rd assessment) and the covariates BMI at baseline, surgery type and smoking status. In order to satisfy a family-wise error rate correction of pFWE<.05, we determined a combined voxel-wise-[p<.001] and cluster-extent-threshold [k ≥ 103] by running 10.000 permutations by Monte Carlo simulations (the estimated smoothness was x/y/z = 10.13/9.86/10.33 mm) using the Neuroelf analysis package (www.neuroelf.net) (Bennett et al., 2009).

2.2.4 Results

Sample characteristics

Demographical, clinical and psychometric data are depicted in Table 2.3.

Table 2.3 Demographic and clinical characteristics of obese study participants that underwent three imaging assessments at To = two weeks prior to surgery, T1 = eight weeks after surgery and T2 = twenty-four weeks after surgery (N = 11).

N = 11 participants with obesity	**Absolute numbers**	**Relative proportions (%)**
Sex (male/female)	3/8	27.3/72.7
Smoking Status (non-smoking/<10 cig. per day/>=10 cig. per day)	7/2/2	63.6/18.2/18.2
	Mean	**SD**
Age (years)	41.18	10.1
Height (m)	1.68	0.1
Weight (kg)	128.78	17.1
BMI	45.40	4.7
BDI (total score)	9.45	4.6

BDI = Beck Depression Inventory, BMI = Body Mass Index

Group-level food cue-induced brain activation

Group-level analyses of brain activation demonstrated significant food cue-induced brain activation (contrast: food > neutral) in parts of the frontal and orbitofrontal cortex, the occipital and parietal gyri, the cuneus, calcarine, the lingual gyrus, as well as the caudate, putamen, thalamus and insula (see Table 2.4). On the other hand, no significant brain activation was detected during presentation of neutral pictures compared to food pictures (contrast: neutral > food). Whole analyses of longitudinal changes in brain responses towards food cues over assessment sessions before and after surgery showed no main effect of time on brain response towards food cues and post-hoc comparisons between separate assessment time points did not surpass the predefined statistical threshold.

Table 2.4 Brain areas depicting higher brain response to visual food cues compared to neutral cues (contrast: food > neutral, combined voxel-wise- [p < .001] and cluster-extent-threshold [k > 103 voxel], corresponding to pFWE < .05).

Side	Lobe	Brain areas	Cluster size (voxel)	MNI coordinates (x, y, z)			t_{max}
R & L	Occipital	Superior, Middle and Inferior Occipital Gyrus, Calcarine, Cuneus, Fusiform Gyrus, Lingual Gyrus	7081	32	-76	-14	21.9
R	Parietal	Inferior Parietal Gyrus, Angular Gyrus	133	32	-68	54	9.6
L	Occipital, Parietal	Superior and Middle Parietal and Occipital Gyrus	275	-24	-60	44	8.9
L		Putamen, Insula	129	-40	-6	10	8.7
L	Parietal	Inferior Parietal Gyrus, Postcentral Gyrus, Supramarginal Gyrus	142	-48	-24	40	8.6
R & L		Anterior and Middle Cingulate Gyrus	176	-8	24	24	7.4
L	Frontal	Middle and Inferior Frontal Gyrus, Orbitofrontal Cortex	130	-44	36	14	7.2
R		Caudate, Thalamus	104	14	-4	12	6.9

Reliability analyses

Food craving ratings

Analyses indicated good reliability of the mean subjective food craving ratings during fMRI across the different assessment sessions (ICC[3,1]=0.611, p = 0.002). Food cues induced higher craving values compared to neutral cues throughout all three assessment sessions. There was a significant reduction in the magnitude of reported food craving over the trial period from baseline (M=45.195, SD=23.443) to T1 (M=18.550, SD=39.917) that remained stable until T2 (M=32.450, SD=25.972, $F_{(2,18)}$ = 4.301, p=0.032).

Jaccard coefficient

Mean Jaccard coefficients for the comparisons of the different time points are displayed in Table 3. Repeated measures ANOVA showed a significant main effect of contrast image category (neutral, food and food>neutral) ($F_{(2,20)}$=83.806 p < 0.001) on the magnitude of the Jaccard indices. Post-hoc analyses demonstrated lower Jaccard coefficients for the difference contrast condition (food>neutral) compared to both constituting conditions (food and neutral, p<0.001). There was no main effect of time on the magnitude of the Jaccard coefficients (i.e. whether we compared to 1st to 2nd or 3rd scanning session, $F_{(2,20)}$=0.152 p = 0.860).

Dice Coefficient

The mean Dice coefficients for the comparisons of the different fMRI sessions are depicted in Table 3. Analyses demonstrated a significant main effect of contrast image category (neutral, food and food>neutral) ($F_{(2,20)}$=77.102 p < 0.001) on the magnitude of the Jaccard indices. Post-hoc analyses demonstrated lower Jaccard coefficients for the difference contrast condition (food>neutral) compared to both constituting conditions (food and neutral, p<0.001). There was no main effect of time on the magnitude of the Jaccard coefficients (i.e. whether we compared to 1st to 2nd or 3rd scanning session, $F_{(2,20)}$=0.208 p = 0.814).

Table 2.5 A] Dice and B] Jaccard coefficients for the three task contrasts (food > neutral, food and neutral), illustrating the proportion of overlapping significant voxels between the different fMRI sessions at To = two weeks prior to surgery, T1 = eight weeks after surgery and T2 = twenty-four weeks after surgery (whole brain threshold of $p < 0.001$ for defining super-threshold activation).

A] Dice Coefficients									
Comparison of	**Session 1 and 2**			**Session 1 and 3**			**Session 2 and 3**		
Contrast	*Food > Neutral*	*Food*	*Neutral*	*Food > Neutral*	*Food*	*Neutral*	*Food > Neutral*	*Food*	*Neutral*
Mean	0.2743***	**0.6763**	**0.7103**	0.2049***	**0.7181**	**0.7260**	0.2218***	**0.6921**	**0.6790**
SD	0.2036	0.2160	0.2067	0.1599	0.0763	0.0762	0.1918	0.2187	0.2106
B] Jaccard Coefficients									
Comparison of Sessions	**Session 1 and 2**			**Session 1 and 3**			**Session 2 and 3**		
Contrast	*Food > Neutral*	*Food*	*Neutral*	*Food > Neutral*	*Food*	*Neutral*	*Food > Neutral*	*Food*	*Neutral*
Mean	0.1744***	**0.5400**	**0.5772**	0.1222***	**0.5651**	**0.5750**	0.1375***	**0.5596**	**0.5406**
SD	0.1443	0.1970	0.1830	0.0996	0.0911	0.0935	0.1316	0.2024	0.1853

SD = standard deviation; *** = significant difference at p < 0.001 between the contrast condition food > neutral and each of the other two conditions (food and neutral)

Intraclass Correlation Coefficient

Comparisons of ICC coefficients between the different fMRI sessions (1st, 2nd, 3rd) indicated that several regions showed good to excellent reliability (i.e. ICC > 0.75) across all sessions (see Figure 1). These regions included the bilateral caudate and left putamen, parts of the right thalamus and middle cingulum, as well as parts of the bilateral inferior, middle and superior occipital gyri (brodmann areas BA 7/17/18/19/39) and parts of the bilateral middle and superior temporal gyri (BA 20/21/22/37) and in addition parts of the bilateral cunei, lingual gyri and calcarine (see Figure 1). These patterns appeared to be relatively stable across all session time points, supporting the stability of the observed findings.

In a second step, we determined the mean ICC for a standard set of n = 120 anatomical regions of interest defined in the aal atlas. As expected, based on the patterns of voxel-wise ICC values (i.e. good to excellent reliability only in parts of the anatomical region), the mean overall ICC for the separate regions did not exceed the voxel-wise values. However, several anatomical regions of interest masks showed good or fair reliability (see supplementary Table S1), specifically the bilateral inferior, middle and superior occipital gyri ROIs showed good overall reliability (>0.6) and the several other regions showed fair reliability (>0.4) Left putamen, bilateral caudate, left amygdala, bilateral lingual gyri, right fusiform gyrus, bilateral calcarine, bilateral cunei, posterior cingulate, right middle temporal gyrus, bilateral middle frontal gyri, right superior medial gyrus, left superior parietal gyrus and angular gyrus. The ICC maps underlying the presented results are provided on Neurovault.org (https://identifiers.org/neurovault.collection:9026).

Brain regions with ICC values > 0.75 (“good“) for contrast food > neutral and N = 11 obese patients

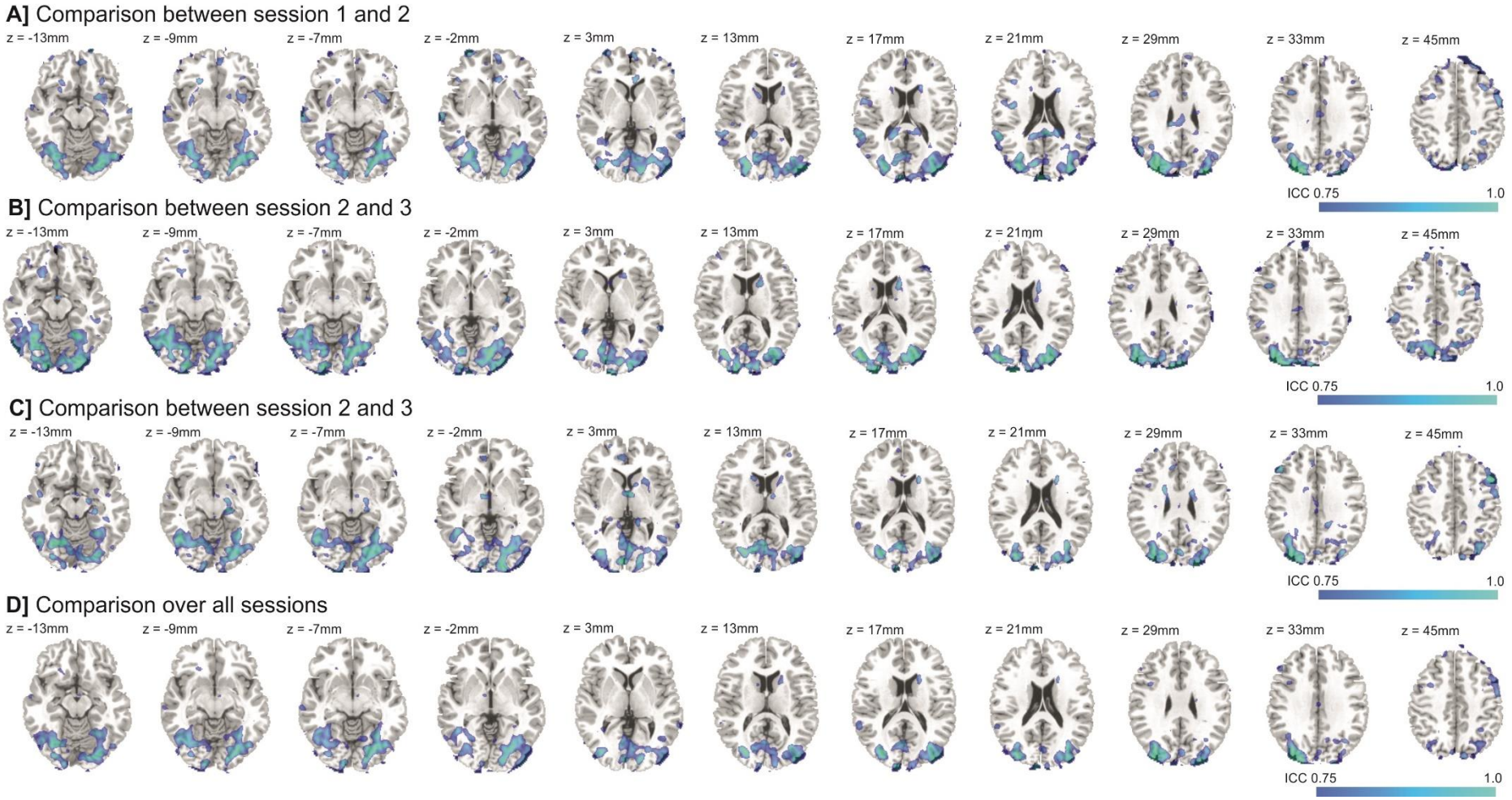


Figure 2.5 Depiction of brain areas that show good to excellent reliability for the difference contrast food-neutral (Intraclass correlation [ICC] > 0.75) for the comparisons between: A] session 1 and 2 (i.e. two weeks prior to surgery and eight weeks after surgery), B] session two and three (i.e. two weeks prior to surgery and twenty-four weeks after surgery), C] session 1 and three and D] over all sessions.

Spearman's rho

We computed spearman's rho coefficients between the food category contrast maps and the neutral contrast maps, in order to assess whether there is a high correlation between the constituting conditions (food and neutral), which would reduce the maximum possible reliability of the difference contrast (food-neutral), due to elimination of shared variance during performing the subtraction. Results demonstrate a substantial correlation between the all three food stimuli category contrast maps and the neutral stimuli contrast maps ($rho_{sweet\text{-}neutral}$ = 0.49, SD = 0.29, R^2=0.24, $rho_{low\text{-}neutral}$ = 0.42, SD = 0.33, R^2=0.17, $rho_{high\text{-}neutral}$ = 0.42, SD = 0.33, R^2=0.17). This indicates that both food and neutral conditions share about 17 to 24% of their variance. A part of this variance is removed by subtracting both conditions, which results in lower reliability of the difference contrast (Infantolino et al., 2018).

Similarity

The analyses of similarity between activation maps for the difference contrast (food > neutral) showed a gradual decrease of within-subject similarity over comparisons between fMRI sessions with increasing time between the respective sessions (i.e. higher within-subject similarity between T0 and T2 that were 10 weeks apart vs. T2 and T3 that were 16 weeks apart). This reflected in lower t-values for the comparisons between within-subject and between-subject similarity for the respective sessions and lower mean similarity values (r_{T0_T1}=0.37, t_{T0_T1}=5.14, p<0.001, r_{T1_T2}=0.32, t_{T1_T2}=3.82, p<0.05, r_{T1_T3}=0.29 t_{T1_T3}=3.01, p<0.05). The difference between within and between-subject similarity is visible in the matrices and cumulative distribution functions for within- and between-subject similarity in Figure 2. The proportion of patients that could be re-identified based on their neural brain activation (i.e. the magnitude of within-subject similarity exceeded all between-subject similarity values). While about 73% of the patients could be re-identified between T0 and T1, this number dropped when comparing longer time periods between T1 to T2 (64%) and T0 to T2 (45%, see Figure 2).

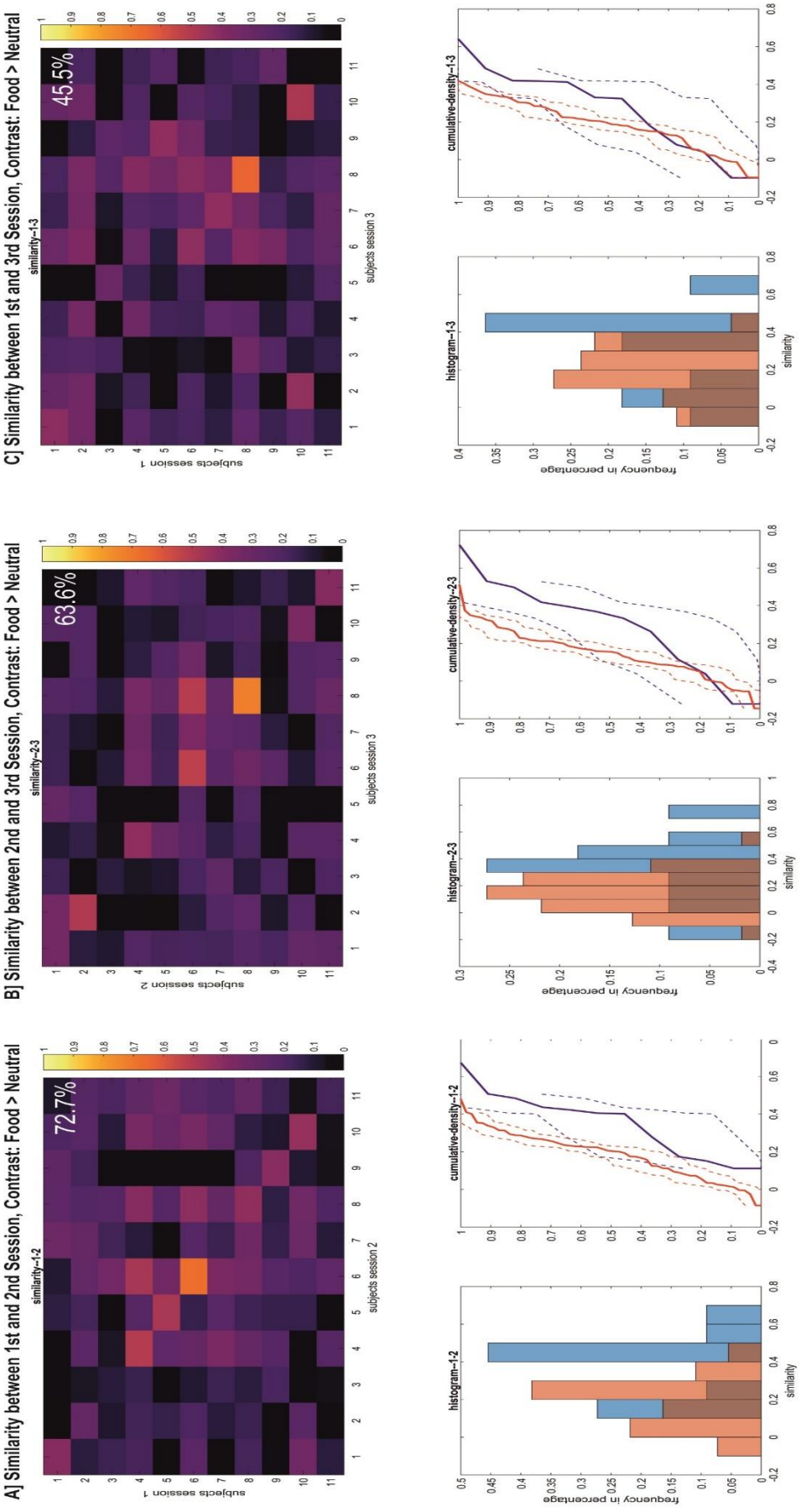

A] Similarity between 1st and 2nd Session, Contrast: Food > Neutral
similarity--1-2
72.7%
subjects session 1
subjects session 2
cumulative-density--1-2
histogram--1-2
frequency in percentage
similarity
B] Similarity between 2nd and 3rd Session, Contrast: Food > Neutral
similarity--2-3
63.6%
subjects session 2
subjects session 3
cumulative-density--2-3
histogram--2-3
frequency in percentage
similarity
C] Similarity between 1st and 3rd Session, Contrast: Food > Neutral
similarity--1-3
45.5%
subjects session 1
subjects session 3
cumulative-density--1-3
histogram--1-3
frequency in percentage
similarity

Figure 2.6 Similarity maps (upper row) and empirical cumulative distribution functions (lower row – red lines: between-subject similarity, blue lines: within-subject similarity) for the contrast food-neutral and comparisons between A] 1st and 2nd fMRI session, B] 2nd and 3rd fMRI session and C] 1st and 3rd fMRI session. The diagonal of each color matrix represents the within-subject similarity values. Re-identification of a subject based on the neural activation map is affirmed 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). Percent values in the upper right of the upper row panels represent the number of individuals that could be identified based on their brain response (i.e. within-subject similarity values exceeded all between-subject similarity values for the respective participant [rows in matrix]).

2.2.5 Discussion

The purpose of this study was to investigate the longitudinal reliability of the different task contrasts an established food cue-reactivity task. ICC values indicated good to excellent reliability of brain activation, captures by the common difference contrast food vs. neutral, in a range of brain areas (i.e. the mesolimbic system with putamen and caudate, as well as parts of the frontal and occipital cortices) over a time period of 26 weeks. In addition, the reliability of food cue-induced brain activation in these brain regions, indexed by the difference contrast food vs. neutral, outperformed the reliability of subjective food craving (i.e. craving during food blocks vs. neutral blocks) that was measured concurrently during fMRI using visual analogue scales. Still, it should be noted that local reliability did not surpass the threshold for good reliability in all areas of the mesocorticolimbic system, which were implicated in processing food cues (Noori et al., 2012). Furthermore, Jaccard and Dice coefficients, which provide estimates for the replicability of significant activation clusters across the whole brain, indicated that only a small proportion of activation could be replicated, when investigating the difference contrast (food > neutral). This stood in sharp contrast to the results for the constituting task contrast conditions food vs. baseline and neutral vs. baseline separately. For these two contrast conditions Jaccard and Dice coefficients showed that more than 50% of the super-threshold clusters could be replicated during the other assessment sessions. This indicates that the global reliability of the common difference contrast food vs. neutral is limited. Several reasons might account for these findings. In previous studies, Infantolino and colleagues (2018) argued that the correlation between the constituting contrast conditions of a difference contrast place a limit on the reliability of the resulting difference measure, because in this case, large proportions of the shared and potentially true variance are eliminated by subtracting both constituting task conditions. The authors sustained their argument with data on the difference contrast between face- and shape-matching trials of a so-called faces paradigm, where the constitution shape and face conditions correlated to 0.97 (Hariri et al., 2002; Infantolino et al., 2018). Other fMRI studies that also computed difference

contrasts as the measure of interest, reported higher reliability of brain activation that was mirrored by an only modest correlation between the constituting conditions (Luking et al., 2017). Current data show a moderate correlation between the food and neutral contrast images with a shared variance of about 24%. This supports the notion that the global reliability of the difference contrast (food vs. neutral) in the current dataset is limited by the correlation between the constituting conditions, which results in an elimination of proportions of the shared variance. The similarity analyses indicated that the capacity to identify individual individuals based on their individual brain activation pattern during the food vs. neutral contrast fades, when time periods between sessions increase. This was an expected finding and suggests that in the case of food cue-reactivity, follow-up fMRI scans should not be scheduled too far apart, when one intends to yield high reliability.

The only other previous study specifically investigating reliability of food cue-reactivity used a pre-selected range of ROIs (insula, putamen, amygdala, orbitofrontal cortex, caudate) and reported overall poor reliability in these ROIs. Several reasons might have accounted for the differences between this and the current study. The study by Sayer et al. (2016) used a different fMRI task design. The number of blocks of neutral and food stimuli per run was markedly lower (i.e. 3 and 3) compared to the task that was used as a basis for current analyses. Fewer data points per subject might however lead to less robust estimates of the individuals "true" mean value, e.g. brain response. Additionally, the study did not investigate voxel-wise reliability, but instead extracted brain activation estimates from predefined regions of interest and focused on the ICC as only an estimate for reliability. The use of the local maxima that were detected in the group level analyses as center of these ROIs, might have biased results. Studies have shown that a robust effect on the group level does not indicate stability or reliability of within-subject effects and might also be influenced by outliers (Infantolino et al., 2018). Hence, the focus on these specific ROIs that only covered a diameter of 3mm around the activation maximum, might have limited the possibility to identify regions with robust reliability. Current atlas-based summaries support the notion that

the areas under investigation, specifically the caudate, putamen and amygdala show at least moderate reliability, when using the fMRI task of the current study.

Multiple studies intended to determine predictors for successful weight loss after bariatric surgery and establish neural "biomarkers" (Holsen et al., 2018; Ness et al., 2014). As reliability is a prerequisite for any measure that could potentially serve as "biomarker", current results could inform future studies and support the notion that neural responses to food cues in a selected range of brain areas might indeed meet the requirements for a potential predictor of treatment outcome.

Strengths and Limitations

We investigated a specific block-design food cue-reactivity task that was used and validated in previous work by our group (Grosshans et al., 2012). Hence, our results may be generalized for food cue-reactivity tasks that incorporate different picture sets or a different task design. Still, the convergence of the different reliability estimates supports the robustness of the findings and the applied methods. We also acknowledge that other methods for the estimation of fMRI reliability exist (e.g. support vector machine learning) and might be informative. We investigated the reliability in a clinical population undergoing surgery. Due to the fact that reliability depends on the population under investigation, we argue that this approach complements the investigation of healthy reference samples, in order to assess the potential of fMRI-based markers for application in clinical populations. Still, the investigation of healthy samples and individuals with obesity are essential, in order to yield robust estimates of reliability of food cue-reactivity without potential bias and reduction in reliability due to surgical intervention or weight loss and improve the overall precision of reliability estimates. We, however, intended to provide a conservative estimate of the reliability of food cue-reactivity, because we acknowledge that statistical control might not be feasible and is also arbitrary to a certain extent (e.g. only controlling variables that show a significant effect of time in a respective trial would lead to differences between trials). This might lead to bias in the estimating of the reliability of food cue-reactivity. It

could be argued that the inclusion of patients without any treatment might be favorable with regards 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 e.g. treatment response, reliability of this putative biomarker should be tested under the very same conditions. It should be noted that reliability estimates, which are based on small datasets, are prone to imprecision, due to large confidence intervals and high impact of single participant data, which also accounts for the presented dataset. The complementary whole brain analyses that compared brain responses towards food cues between the different assessment session did not yield significance, when applying a stringent whole-brain correction for multiple testing. This result is unexpected and contrasts previous studies that showed longitudinal changes in brain response from before to after surgery (Li et al., 2019; Zoon et al., 2018). The lack of significant main effects of time on brain response might relate to a limited power and a stringent whole brain threshold (e.g. previous studies applied regions of interest analyses), resulting from the small dataset. However, power analyses indicated that analyses comparing different time points yielded sufficient power (see Supplementary Figure S1). Additionally, several significant findings were derived from studies applying more liberal regions of interest analyses. Overall, the lack of substantial time effects on the extent of food cue-induced brain response in the current dataset support the notion that reliability estimates were not substantially biased by surgical intervention.

2.2.6 Conclusion

We could show excellent local longitudinal reliability in a range of brain areas of the reward (e.g. caudate, putamen) and food-cue processing networks (e.g. occipital and frontal cortices) in participants with obesity from two weeks before, to 24 weeks after surgery. The reliability of food cue-reactivity in these areas outperformed to reliability of subjective craving measures that were measured concurrently. Our results suggest that fMRI-based measures might indeed be suitable to monitor and predict treatment outcome in participants with obesity undergoing bariatric surgery.

2.2.7 Supplements

Supplementary Tables

Supplementary Table S 2.5 Atlas-based mean intraclass correlation (ICC) values for the N=120 anatomical regions specified in the automated anatomical labeling (aal) atlas (contrast: food > neutral stimuli, comparisons across sessions 1 to 3). Regions exceeding a mean ICC value of 0.4, corresponding to a moderate reliability, are marked in bold font.

	Region (AAL)	Mean ICC value
1.	**Amygdala_L**	**0.40763759**
2.	Amygdala_R	0.36786118
3.	**Angular_L**	**0.4153282**
4.	**Angular_R**	**0.48685384**
5.	**Calcarine_L**	**0.53597925**
6.	**Calcarine_R**	**0.57968044**
7.	**Caudate_L**	**0.48102494**
8.	**Caudate_R**	**0.44163647**
9.	Cerebelum_10_L	0.27364602
10.	Cerebelum_10_R	0.09230089
11.	Cerebelum_3_L	0.06165687
12.	Cerebelum_3_R	0.17436859
13.	Cerebelum_4_5_L	0.15992892
14.	Cerebelum_4_5_R	0.21595067
15.	Cerebelum_6_L	0.30154894
16.	Cerebelum_6_R	0.32449259
17.	Cerebelum_7b_L	0.00027952
18.	Cerebelum_7b_R	0.15759062
19.	Cerebelum_8_L	-0.02479433
20.	Cerebelum_8_R	0.10007577
21.	Cerebelum_9_L	0.12451792
22.	Cerebelum_9_R	0.06073853

23.	Cerebelum_Crus1_L	0.34667738
24.	Cerebelum_Crus1_R	0.34672601
25.	Cerebelum_Crus2_L	0.16982958
26.	Cerebelum_Crus2_R	0.26460454
27.	Cingulate_Ant_L	0.31521474
28.	Cingulate_Ant_R	0.29600421
29.	Cingulate_Mid_L	0.27712185
30.	Cingulate_Mid_R	0.27994037
31.	**Cingulate_Post_L**	**0.50434102**
32.	**Cingulate_Post_R**	**0.49553011**
33.	**Cuneus_L**	**0.56512844**
34.	**Cuneus_R**	**0.5228328**
35.	Frontal_Inf_Oper_L	0.14184759
36.	Frontal_Inf_Oper_R	0.21942506
37.	Frontal_Inf_Orb_2_L	0.27682627
38.	Frontal_Inf_Orb_2_R	0.21755721
39.	Frontal_Inf_Tri_L	0.30462771
40.	Frontal_Inf_Tri_R	0.24102302
41.	Frontal_Med_Orb_L	0.25330428
42.	Frontal_Med_Orb_R	0.25289913
43.	**Frontal_Mid_2_L**	**0.41863018**
44.	**Frontal_Mid_2_R**	**0.4209279**
45.	Frontal_Sup_2_L	0.32470793
46.	Frontal_Sup_2_R	0.32815861
47.	**Frontal_Sup_Medial_L**	**0.45450867**
48.	**Frontal_Sup_Medial_R**	**0.40819146**
49.	**Fusiform_L**	**0.54252807**
50.	**Fusiform_R**	**0.58699878**
51.	Heschl_L	0.05446871
52.	Heschl_R	0.20525357
53.	Hippocampus_L	0.30091927

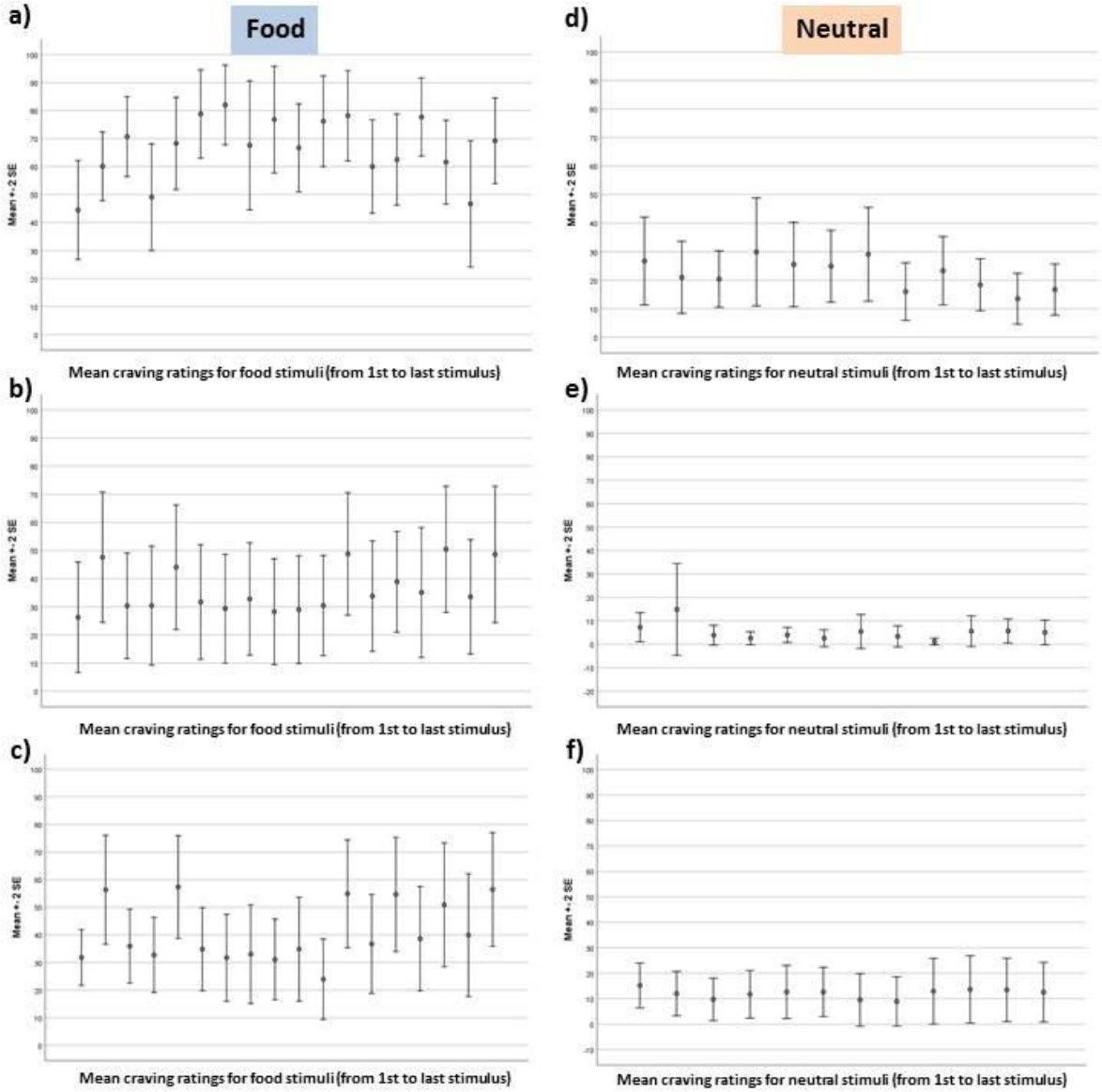
54.	Hippocampus_R	0.18147327
55.	Insula_L	0.20153209
56.	Insula_R	0.1647087
57.	**Lingual_L**	**0.6225418**
58.	**Lingual_R**	**0.57012037**
59.	**Occipital_Inf_L**	**0.68905961**
60.	**Occipital_Inf_R**	**0.63069714**
61.	**Occipital_Mid_L**	**0.70694477**
62.	**Occipital_Mid_R**	**0.73225597**
63.	**Occipital_Sup_L**	**0.6863284**
64.	**Occipital_Sup_R**	**0.64287727**
65.	OFCant_L	0.20520715
66.	OFCant_R	0.26953573
67.	OFClat_L	0.11979768
68.	OFClat_R	0.31150854
69.	OFCmed_L	0.10423343
70.	OFCmed_R	0.18515158
71.	OFCpost_L	0.30889935
72.	OFCpost_R	0.24117885
73.	Olfactory_L	0.24754212
74.	Olfactory_R	0.26154753
75.	Pallidum_L	0.27670088
76.	Pallidum_R	0.21066503
77.	Paracentral_Lobule_L	0.14723272
78.	Paracentral_Lobule_R	0.06484738
79.	ParaHippocampal_L	0.30799088
80.	ParaHippocampal_R	0.28349796
81.	Parietal_Inf_L	0.26225495
82.	Parietal_Inf_R	0.22906141
83.	**Parietal_Sup_L**	**0.41622173**
84.	Parietal_Sup_R	0.3988176

85.	Postcentral_L	0.16078831
86.	Postcentral_R	0.08968218
87.	Precentral_L	0.39168091
88.	Precentral_R	0.18921363
89.	Precuneus_L	0.36465171
90.	Precuneus_R	0.35655186
91.	**Putamen_L**	**0.45442722**
92.	Putamen_R	0.30970413
93.	Rectus_L	0.17306424
94.	Rectus_R	0.14920027
95.	Rolandic_Oper_L	0.13431982
96.	Rolandic_Oper_R	0.10625991
97.	Supp_Motor_Area_L	0.26899723
98.	Supp_Motor_Area_R	0.18813331
99.	SupraMarginal_L	0.05553424
100.	SupraMarginal_R	0.36478604
101.	Temporal_Inf_L	0.35265871
102.	Temporal_Inf_R	0.34174827
103.	Temporal_Mid_L	0.2828941
104.	**Temporal_Mid_R**	**0.46616954**
105.	Temporal_Pole_Mid_L	0.17525254
106.	Temporal_Pole_Mid_R	0.23368946
107.	Temporal_Pole_Sup_L	0.2615089
108.	Temporal_Pole_Sup_R	0.09053107
109.	Temporal_Sup_L	0.18427253
110.	Temporal_Sup_R	0.34549048
111.	Thalamus_L	0.23469258
112.	Thalamus_R	0.23898485
113.	Vermis_1_2	0.19670743
114.	Vermis_10	0.25951009
115.	Vermis_3	0.14516806

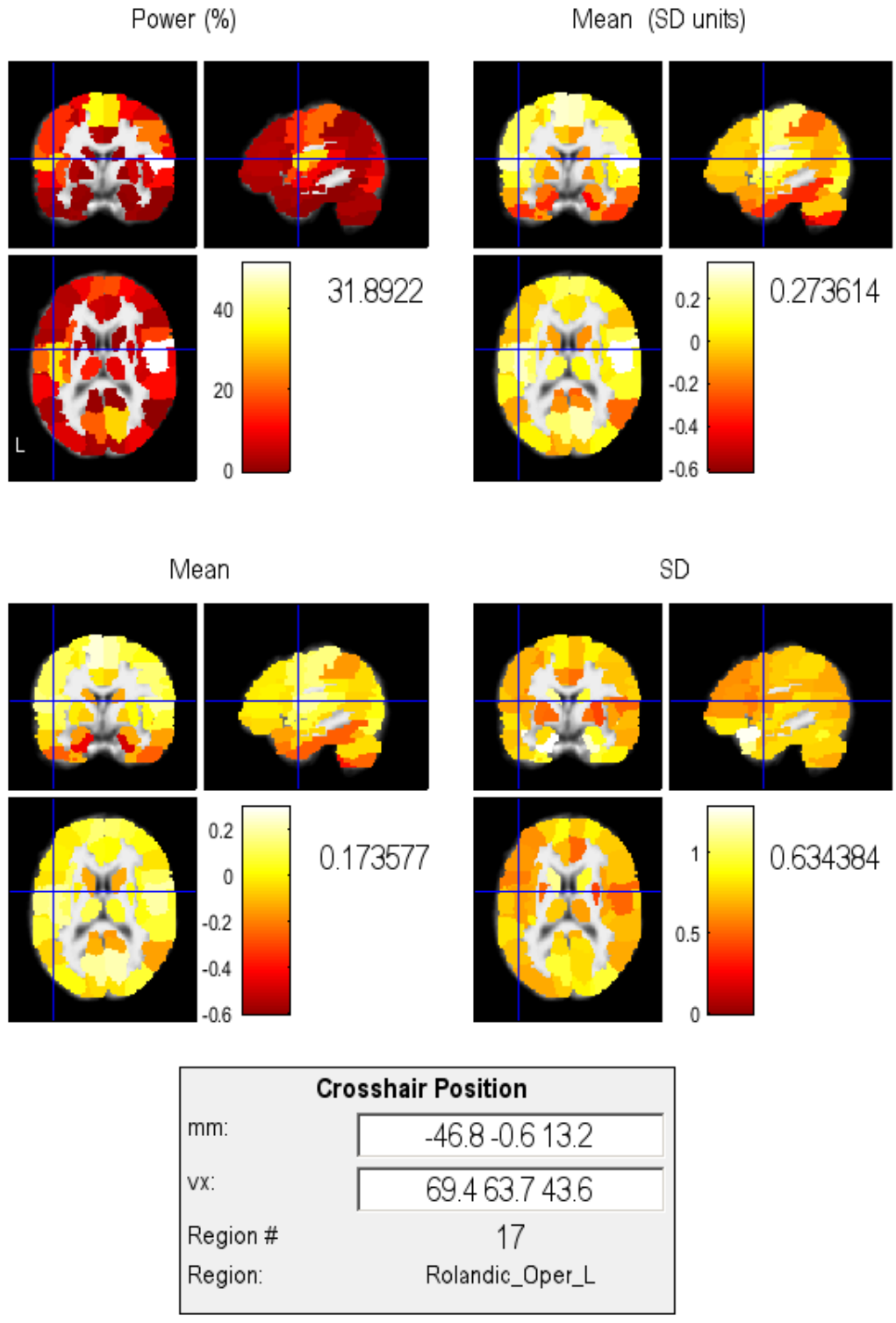
116.	Vermis_4_5	-0.03299619
117.	Vermis_6	0.12162924
118.	Vermis_7	-0.00733943
119.	Vermis_8	-0.18407945
120.	Vermis_9	-0.06657724

Supplementary Figures

Supplementary Figure S 2.5 Depiction of the course of mean craving ratings for food stimuli for the three assessment sessions, for a) the 1st, b) the 2nd and c) the 3rd assessment. Depiction of the course of mean craving ratings for neutral stimuli for the three assessment sessions, for d) the 1st, e) the 2nd and f) the 3rd assessment (Mean + 2 SE).



Supplementary Figure S 2.6 Depiction of power estimates and mean effects for different brain regions defined by the automated anatomical labeling (aal) atlas, which were computed for a pairwise comparison between 1st and 3rd assessment sessions, based on the dataset of N=11 participants, using the FMRIpower software toolbox for SPM (https://www.nitrc.org/projects/fmripower/).



2.3 Study 3 – Reliability of the fMRI-based assessment of self-evaluation in individuals with internet gaming disorder[3]

2.3.1 Abstract

The self-concept—defined as the cognitive representation of beliefs about oneself—determines how individuals view themselves, others, and their actions. A negative self-concept can drive gaming use and internet gaming disorder (IGD). The assessment of the neural correlates of self-evaluation gained popularity to assess the self-concept in individuals with IGD. This attempt, however, seems to critically depend on the reliability of the investigated task-fMRI brain activation. As first study to date, we assessed test-retest reliability of an fMRI self-evaluation task.

Test-retest reliability of neural brain activation between two separate fMRI sessions (approximately 12 months apart) was investigated in N=29 healthy participants and N=11 individuals with pathological internet gaming. We computed reliability estimates for the different task contrasts (self, a familiar, and an unknown person) and the contrast (self > familiar and unknown person).

Data indicated good test-retest reliability of brain activation, captured by the "self", "familiar person", and "unknown person" contrasts, in a large network of brain regions in the whole sample (N=40) and when considering both experimental groups separately. In contrast to that, only a small set of brain regions showed moderate to good reliability, when investigating the contrasts ("self > familiar and unknown person"). The lower reliability of the contrast can be attributed to the fact that the constituting contrast conditions were highly correlated.

[3] Bach P, Hill H, Reinhard I, Gädeke T, Kiefer F, Leménager T. Reliability of the fMRI-based assessment of self-evaluation in individuals with internet gaming disorder. Eur Arch Psychiatry Clin Neurosci. 2021 Jul 17. doi: 10.1007/s00406-021-01307-2. Epub ahead of print. PMID: 34275007.

Future research on self-evaluation should be cautioned by the findings of substantial local reliability differences across the brain and employ methods to overcome these limitations.

2.3.2 Introduction

Converging lines of evidence implicate deficits in an individual's self-concept as a relevant factor in the development of internet gaming disorders (IGD). Due to their social and concomitantly anonymous characteristics, reports portray internet games as a tempting pastime to compensate a negative self-concept. The self-concept can be regarded as a relatively stable cognitive representation (i.e., knowledge system and beliefs) about the own person (Rammsayer & Weber, 2016; Swarm Jr, 1983). It is defined as the subjective evaluation of one's own physical appearance (physical self-concept); social competences (social self-concept); the capability to recognize, express, and regulate one's feelings (emotional self-concept); and skills for reaching academic goals (academic self-concept) (Mummendey, 2006). Mummendey (2006) assumes that the self-concept evolves from comparisons between the subjective view of oneself and the ideal self (i.e., how one would like to be). The ideal self is mainly influenced by an individual's environment, such as family members, peer groups, society, and media. A negative self-concept—defined as a high discrepancy in the evaluation between the subjective self and ideal self—is often reported to be associated with gaming and other internet related disorders (Leménager et al., 2018; Lemmens et al., 2015; Wartberg et al., 2017; Wartberg et al., 2019). A recent review on self-esteem and self-concept in gaming disorders, reported stable associations between gaming disorders and negative physical and academic self-concept domains (Lemenager et al., 2020). Functional imaging studies have tried to identify the neural correlates of the self-concept by applying self-referential and self-recognition paradigms. During self-referential tasks, participants are asked to evaluate their personality traits, physical appearance, preferences, or thoughts. Resulting neural activation patterns are compared to those that emerge during the evaluation of one's ideal self, a close friend, a famous person, or a foreign person (Fossati et al., 2004; Johnson et al., 2002; Lou et al., 2004).

The neural activations during self-referential paradigms are regarded as a functional network underlying different self-concept aspects (e.g., reflecting on one's own social competencies can be regarded as part of the social self-concept). Self-recognition paradigms, in which participants see pictures of their own face or body relative to faces or bodies of others (Northoff et al., 2006), involve unconscious comparisons between one's own physical appearance and that of others. It can be assumed that these comparisons mirror neurobiological correlates of the physical self-concept.

A meta-analysis of Hu et al. (Hu et al., 2016) compared neural correlates of self-face recognition and self-referential paradigms in healthy participants to identify distinct and common neural regions underlying self-referential and self-recognition. Processing one's own face relative to the face of another person, induced activation in the right inferior frontal gyrus (IFG); the bilateral fusiform gyrus; the inferior temporal gyrus; the bilateral insula; the right postcentral and supramarginal gyrus (SMG); the anterior cingulate (ACC); and the right superior occipital and angular gyrus. The meta-analysis also indicated that, across studies, self-referential paradigms induced brain activation in the bilateral ACC; the middle frontal gyrus and the superior temporal gyrus; the precuneus as well as the left inferior parietal gyrus. The conjunction analysis of both tasks revealed shared activation in the right ACC and in the left insula and IFG (Hu et al., 2016). The results of the meta-analysis also demonstrated that self-referential and self-recognition tasks induce activation in regions of the temporoparietal junction (TPJ), extending over several cortical areas, including posterior portions of the superior temporal gyrus and adjacent parietal regions in the supramarginal and angular gyri (Hu et al., 2016). Studies in addicted gamers revealed increased activation in the right inferior parietal lobule (Kim et al., 2018) and a decrease in the right inferior frontal gyrus (Choi et al., 2018) during self-reflection. Choi et al. (2018) assessed the self-concept in addicted adolescent gamers compared to healthy controls during self-referential tasks. The authors observed a decrease in the inferior frontal gyrus in the addicted group, indicating that addicted gamers might find it more difficult to retrieve information regarding their self-concept. Furthermore, Kim et al. (2018) found an increase in activation in the inferior parietal lobule in the addicted group during self vs.

ideal self-reflection. The authors interpreted their findings as an increased identification with the real self, as compared to the ideal self, in individuals with gaming disorders. The investigation of self-concept-related characteristics via self-evaluation tasks seems to be a promising approach to further elucidate the neurobiological basis of gaming disorders; however, currently there is no data on the reliability of fMRI tasks that assess self-concept in individuals with IGD. Reliability of an fMRI task, however, is an important prerequisite for capturing individual neural correlates of the self-concept and for establishing associations between neural processes and behavior. Furthermore, it is used to predict future behavior in gaming disorders, based on neural activation patterns. Elliot and colleagues (2019) conducted a meta-analysis of fMRI studies and pointed out that the overall reliability of fMRI tasks across different task categories, designs, and study groups was low (Elliott et al., 2020). A study by Infantolino et al. (2018) also indicated that low reliability of fMRI tasks might result when fMRI task contrasts are computed by subtracting two correlated task conditions, even when the constituting task conditions show excellent reliability. This is because much of the shared "true" variance is removed when subtracting two task conditions from one another, while the error variance is summed (Infantolino et al., 2018; Peter et al., 1993). Hence, we set out to assess the reliability of neural responses during a video-based fMRI paradigm assessing neural correlates of physical, social, and emotional aspects of the self-concept in young adults. The video paradigm combines self-referential and self-recognition aspects. During the fMRI session, video clips of the participant, a close friend, and an unknown person are presented. The protagonists in the videos introduce themselves and talk about topics related to the self-concept, such as their positive personality traits (emotional and social self-concept); their expectations of others (social self-concept); as well as their future goals (academic self-concept). We assume that the comparison of neural activation between the self and other conditions mirrors neural correlates of the self-concept. To our knowledge this is the first paradigm measuring self-concept-related aspects by combining self-referential and self-recognition paradigms. We tested this paradigm in the framework of a one-year

longitudinal study in a sample of healthy participants as well as individuals with pathological internet game use.

2.3.3 Methods

Study sample and patient subgroups

A total of N=40 male individuals (n=11 pathological [problematic and addicted] gamers and n=29 controls) were included in the current analyses. Initially, N=83 participants enrolled in the study and completed baseline assessment. Of those, N=40 returned for a second assessment after 12 months. N=40 participants provided complete datasets. Participants were recruited between March 2016 and June 2019 (trial registration: DRKS 00009439). All procedures were carried out in accordance with the Declaration of Helsinki. The local ethics committee (application number 2014-602N-MA) approved the study procedures and all participants provided informed written consent. Individuals were recruited via advertisement and outpatient care for pathological gamers in the Central Institute of Mental Health, Mannheim, Germany. Between the first (T1) and second assessment (T2), participants did not receive any specific intervention. The average time span between T1 and T2 was 396 days (SD=67). Abstinence from substance use was monitored through drug urine screening at each assessment.

Participants were required to be aged between 18 and 27 years and had to be right-handed. Pathological gamers were excluded if they met any of the following exclusion criteria: (i) comorbid axis I disorders in the preceding year aside from nicotine-dependence and internet gaming disorder, assessed using the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID) (H. U. Wittchen et al., 1997) and the Assessment of Internet and Computer Game Addiction (AICA) (Wölfling et al., 2012); (ii) treatment with psychotropic or anticonvulsive medications; (iii) severe neurological or physiological disease (such as, but not limited to stroke, aneurysm, dementia, epilepsy, liver cirrhosis); (iv) negative urine drug test on the day of assessment; or (v) contraindications for MRI scans (i.e., pace-makers, metal implants, tattoos).

Assessment

Participants underwent two assessment sessions, both including psychometric measures and fMRI. All participants completed questionnaires on (and after) the assessment day, including the Rosenberg Self-Esteem Scale (Rosenberg, 1965), the Scale for Social Anxiety and Social Competence Deficits (SASKO; (Kolbeck & Maß, 2009)), the Emotional Competence Questionnaire (Rindermann, 2009) and the Empathy Quotient (Baron-Cohen & Wheelwright, 2004). Diagnosis of internet gaming addiction as well as problematic usage was evaluated with the Assessment of Internet and Computer game Addiction-Checklist (AICA-C>13 for addictive usage and AICA-S >6 and <13 for problematic usage; (Wölfling et al., 2012)). After the first assessment, participants underwent interviews and filled out questionnaires every three months. After 12 months, participants were assessed via fMRI once again. Before the second scan, the exclusion criteria were reconfirmed. Participants were excluded if they had developed a comorbid axis I disorder (other than nicotine-dependence and internet gaming disorder, in the preceding year); if they underwent treatment with psychotropic or anticonvulsive medications; or if they had suffered severe neurological or physiological disease in the preceding 12 months.

fMRI self-evaluation task

The paradigm comprised video clips of the participant themselves, an age-matched familiar person, and an unknown person. The task was programmed with the software Presentation Version 16.3 (Neurobehavioral Systems, Albany, Calif., USA). During a video session, participants and their close friend were asked to introduce themselves and talk about different topics related to their person. Four videos of each condition (self, a familiar, and an unknown person) comprised the following topics: (1) personal introduction (instruction: "Introduce yourself: name, age, family, etc."); (2) positive character traits (instruction: "What are your personal strengths and hobbies?"); (3) personal values and expectations of other people (instruction: "What is important to you concerning your fellow humans?"); and (4) future goals (instruction: "Where do

you see yourself in five years from now, what did you achieve?"). The videos, with duration of 15 seconds each, were recorded in advance with a Panasonic high-definition video camera (Type HC-V707) and converted using VSDS Free Video Editor software (Version 3).

The fMRI paradigm was conducted in a block design. Each paradigm block consisted of a video clip regarding one topic from one specific condition. All blocks were presented in a randomized order. Every participant watched 12 video clips in total (three conditions comprising four videos each). Each video clip was followed by a fixation cross (two seconds) and a distractor (calculation task with a maximum duration of 13 seconds), where participants had to move the cursor to select an answer. Then, another fixation cross appeared before the subsequent video clip began. The distractor was used to create distance between the previous videos' content. The total paradigm took a minimum of four and a maximum of 8 minutes, depending on how fast the participants solved the calculation task.

MRI acquisition

MRI data acquisition was performed on a 3.0 Tesla MR scanner (SIEMENS MAGNETOM Trio) with a standard multi-channel receiver head coil (12-channel). During the functional self-concept MRI task, 205 volumes were acquired by applying a T2*-weighted echo-planar imaging (EPI) sequence [repetition time (TR) = 2410 ms, echo time (TE) = 25 ms, flip angle (FA) = 80°, field of view (FOV) = 192 mm × 192 mm, matrix size 64 × 64, 42 slices, slice thickness = 2.00 mm, distance factor = 50%, and voxel size = 3 × 3 × 2 mm]. Three-dimensional T1-weighted structural images (Magnetization Prepared Rapid Acquisition Gradient Echo, MPRAGE) were collected over 8 minutes. The T1-weighted anatomical scans comprised 192 sagittal slices (flip angle: 9˚; repetition time: 2.3 ms; echo time: 3.03 ms; field of view 256 × 256; voxel size, 1 mm x 1 mm x 1 mm). The automated Siemens Multi-Angle Projection (MAP) Shim corrected magnetic field inhomogeneity. Presentation software (Version 16.3, Neurobehavioral Systems, Inc., Albany, CA, USA) was used for both the registration of scanner triggers and the recording of behavioral responses. All participants viewed the video

clips through a tilted mirror placed above their heads. During the assessment, the test persons wore foam ear plugs and headphones. Prior to the assessment, participants underwent a hearing test to adjust the sound of the video clips if necessary. After completion of the scan, participants were asked to rate the sound quality of the videos on a scale from 0 to 10. One patient, who rated the sound quality under 7, was excluded from the analyses.

fMRI pre-processing and statistical analyses

The functional images were pre-processed according to standard procedures implemented in the statistical parametric mapping software for Matlab (SPM, Wellcome Department of Cognitive Neurology, London, UK) version 12. The first five scans of every measurement were discarded to avoid artifacts due to magnetic saturation. We conducted slice time correction, followed by spatial realignment and unwarping. A phase map correction was applied to correct geometric distortions, using a voxel displacement map that was computed from a gray field mapping sequence using the VDM utility in SPM12. Movement correction was conducted using standard SPM12 parameters and images were normalized to the standard tissue probability template provided in SPM12. Smoothing was conducted using an isotropic Gaussian kernel for group analysis (8 mm Full Width at Half Maximum). The following procedures were carried out to assess the quality of pre-processed functional MRI data. Motion correction and realignment parameters, as well as results from the normalization procedure, were assessed by two independent trained members of the study team. Datasets of participants were excluded if the spatial realignment or movement correction parameters indicated excessive motion (>3 degrees of rotation or >3 mm movement in any axis) or if visual inspection indicated poor fitting to the standard TPM template. The first-level statistics were computed for each participant, modelling the different experimental conditions: (i) self, (ii) familiar person, (iii) unknown person, and (iv) distractor task in a generalized linear model including six motion parameters as covariates. The general view of the self-concept is that of a stable cognitive representation (i.e., knowledge system and beliefs) about one's subjective self in comparison to an ideal self, the latter of which is formed by the environment. In line with this view, the

neural correlates of the self-concept were operationalized by subtracting brain activation during the presentation of videos of oneself from the brain activation during the presentation of videos of familiar and unknown persons (i.e., self > familiar person and unknown person). Thus, apart from the contrast images for (i) self vs. implicit baseline; (ii) familiar person vs. implicit baseline; (iii) unknown person vs. implicit baseline; (iv) distractor condition vs. implicit baseline; and (v) self vs. familiar and unknown person. The contrast between self and familiar person + unknown person was computed using the contrast weights [2 -1 -1 0].

Previous studies suggested that difference measures suffer from low inherent reliability when the constituting conditions are correlated (Infantolino et al., 2018). Hence, we also estimated reliability separately for the "self", "familiar other", and "unknown other" contrast conditions.

Analyses of self-concept-related measures

We tested the stability of self-concept measures (i.e., with SASKO, the Emotional Competence Questionnaire, the Rosenberg Self-Esteem Scale, and the Empathy Quotient) by assessing differences between the first and second experimental session (t-tests for dependent samples) for pathological (problematic and addicted) gamers and healthy controls separately. Furthermore, we assessed test-retest reliability of self-concept measures by computing the intraclass correlation coefficient between the first and second session.

Analyses of group-level fMRI activation

On a group level, imaging data were analyzed using full factorial models with the factor time (first and second scan) to assess the congruence of task effects on the group level brain activation over time. This was accomplished by determining brain areas that show higher brain activation in response to viewing videos of the own person compared to brain activation when viewing videos of familiar and foreign persons (contrast: "self > familiar and unknown person"). In addition, group-level brain activation patterns were analyzed for the constituting task conditions separately (i.e., responses to videos of the "self"); a familiar person (contrast: "familiar person"); and an

unknown person (contrast: "unknown person") at each time point. We applied a whole-brain family-wise error rate correction of $p_{FWE}<.05$ at the cluster level to correct for multiple comparisons.

Reliability measures

To assess longitudinal test-retest reliability of the self-evaluation fMRI task, we computed global and local measures of reliability. All reliability analyses were conducted using the fmreli toolbox for SPM12 (Frohner et al., 2019). Individual contrast images of the different task conditions served as input for the reliability analyses. Dice and Jaccard coefficients were analyzed within the framework of an ANOVA with the contrast condition set as four-level within-subject factor (i. self; ii. familiar person; iii. unknown person; iv. self > familiar + unknown person) and the experimental group set as two-level between-subject factor (i. healthy individuals; ii. IGD).

Intraclass correlation

Voxel-wise reliability of each contrast condition was estimated by computing the intraclass correlation coefficient (ICC) between the first and second assessment points. The ICC tests whether the magnitude of brain activation in each voxel is stable between the first and the second fMRI scan. Fleiss (1986) proposed that ICCs lower than 0.4 indicate poor reliability; ICCs between 0.4 and 0.6 indicate fair reliability; ICCs between 0.6 and 0.75 indicate good reliability; and ICCs with values higher than 0.75 indicate good to excellent reliability (Fleiss, 1986). The ICC sets within-subject variance (σ^2_{within}) in relation to between-subject variance ($\sigma^2_{between}$). The ICC(3,1)-type was proposed as being the most appropriate for assessing single site longitudinal fMRI datasets (Ombao et al., 2016). Hence, we used the ICC(3,1)-type (Shrout & Fleiss, 1979), defined as:

$$ICC = \frac{(\sigma^2{}_{\text{between}} - \sigma^2{}_{\text{within}})}{(\sigma^2{}_{\text{between}} + \sigma^2{}_{\text{within}})}$$

ICC values were computed for the contrasts of "self", "familiar person", "unknown person", and the contrasts "self > familiar and unknown person". We computed ICCs for every brain voxel and generated thresholded ICC brain maps to identify brain areas

that show good (ICC > 0.6) and good to excellent (ICC > 0.75) reliability. Furthermore, we computed additional atlas-based mean ICC values for a standard set of anatomical brain regions (see below).

Similarity

Similarities in the fMRI activation maps from the first and second scans were determined. The analysis captures the resemblance of two brain activation maps based on the alignment of high vs. low brain activation values across the brain. The authors of the fmreli toolbox propose that this method could be used to quantify within-subject and between-subject similarities of brain activation without requiring an a priori (and potentially arbitrary) statistical threshold. A high within-subject similarity supports the notion that individuals can be re-identified based on their neural brain activation patterns. The resulting coefficients are correlation coefficients that range from a "perfect" negative relationship (-1.00) to a "perfect" positive relationship (1.00). In the past, studies have suggested that subjects can be successfully identified based on their neural activation pattern if the within-subject similarity exceeds all between-subject association coefficients of the same participant (Finn et al., 2015; Frohner et al., 2019). The similarity analyses, therefore, complement the computation of the ICC, which allow inferences on a group level, providing additional information on the stability and resemblance of brain activation at an individual participant level.

Pearson's correlation

We computed the mean voxel-wise Pearson's correlation coefficients between the "self", "familiar other", and "unknown other" contrast conditions using the procedures provided in the fmreli toolbox. This step was taken to assess the correlation between the different task condition contrasts. This is important due to the fact that the reliability of a contrast between two conditions is limited in the case of high correlation between the activation patterns of the constituting contrast conditions.

Jaccard and Dice Coefficients

The modified Jaccard coefficient is a commonly used measure in fMRI reliability studies. It can be interpreted as the percentage of overlapping significant voxels above a predefined threshold (e.g., p < 0.001) within all significant voxels. The Jaccard coefficient is defined as the ratio of intersection between the number of three-dimensional image voxels, which were found to be activated in the first fMRI assessment (A) and the replication (B), divided by the size of the union of the voxel sets of A and B (Jaccard, 1902; Maitra, 2010).

$$Jaccard(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Another measure of global reliability or overlap between super-threshold voxels is the Dice coefficient. It is calculated as the number of super-threshold voxels that overlap between sessions A and B (see above) divided by the average number of significant voxels across sessions A and B (see above):

$$Dice(A,B) = \frac{2|A \cap B|}{|A| + |B|}$$

Both, Jaccard and Dice coefficients range from no overlap (0) to perfect overlap (1) between super-threshold voxels; however, currently there is no consensus on specific values or cut-offs that would differentiate between "poor" and "good" values (Bennett & Miller, 2010). In accordance with previous studies, the current analyses used a threshold of p < 0.001. Jaccard and Dice coefficients were determined for every patient by comparing the baseline and the second fMRI results for the different contrast images. Resulting values were exported into the IBM SPSS statistics software (version 25.0) and effects of contrast conditions were tested using a repeated measures analysis of variance model with contrast condition as within-subject factor.

Atlas- and ROI-based summary measures

To facilitate the assessment of local differences in reliability, we computed the mean ICC for N=116 anatomical regions, specified in the Automatic Anatomic Labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002). ICC values were extracted from the ROIs

using the data extraction routine of the MarsBar software package (http://mars-bar.sourceforge.net/); then, these data were exported into the IBM SPSS statistics software (version 25.0) for further analyses.

2.3.4 Results

Reliability of psychometric self-concept-related measures

Demographic, gaming, and self-concept-related psychometric data for both time points of N=40 participants are shown in Table 2.6. The mean period between the two fMRI scans was 396 days (SD=67). As expected, pathological gamers showed significantly higher AICA_30 and AICA_lifetime scores than healthy controls at T1 and T2. The AICA_30 score showed a significant decrease from T1 to T2 for the pathological gaming group but not for the healthy control group. Regarding between-group differences, pathological gamers rated their social anxiety of feeling rejected higher, compared to the control groups in T2. Both subgroups did not differ in other self-concept-related measures. Furthermore, the majority of self-concept-related measures did not show a significant change over time, which indicates stability (see Table 2.6). Only the subscale assessing emotional regulation (EKF-RE) showed a significant increase from T1 to T2 in pathological gamers. The stability of self-concept-related measures was further supported by high test-retest reliability estimates (see Table 2.7).

Table 2.6 Sample description - Differences in demographic, gaming, as well as self-concept-related characteristics between pathological (problematic and addicted) gamers and healthy controls at both time points (t-tests for independent samples) and between the time points (t-tests for dependent samples).

Between Group Comparison T1 – 1st Assessment					
	Total score	Pathological gamers (n=11)	Healthy controls (n=29)	t-value	p-value
Age (SD)	21.23 (2.51)	21.45 (3.05)	21.14 (2.33)	-.353	.726
AICA_30 (SD)	5.15 (5.89)	13.09 (5.55)	2.14 (1.79)	-.641	.007
AICA_lifetime (SD)	13.70 (8.36)	22.27 (5.73)	10.45 (6.77)	-5.125	<.001
Years of education (SD)	13.92 (1.90)	14.09 (1.92)	13.86 (1.92)	-.336	.739
SASKO: Speaking (SD)	8.40 (5.92)	10.64 (6.91)	7.55 (5.38)	-1.33	.202
SASKO: Rejection (SD)	7.95 (5.27)	10.36 (5.83)	7.03 (4.83)	-1.839	.074
SASKO: Interaction (SD)	6.10 (4.75)	7.36 (7.00)	5.62 (3.62)	-1.03	.307
SASKO: Information (SD)	6.08 (3.82)	8.36 (4.80)	5.20 (3.04)	-2.032	.063
SASKO: Loneliness (SD)	2.15 (2.50)	2.73 (2.76)	1.93 (2.40)	-.898	.375
EKF-EE (SD)	56.90 (8.27)	56.73 (9.81)	56.97 (10.18)	.067	.947
EKF-EA (SD)	62.552 (8.27))	61.18 (10.57)	63.03 (7.38)	.628	.534
EKF-RE (SD)	48.25 (6.89	47.18 (8.11)	48.66 (6.48)	.599	.553
EKF-EX (SD)	51.25 (11.18)	50.73 (13.88)	51.44 (10.25)	.180	.858
Rosenberg Selfworth (SD)	22.90 (4.87)	21.36 (6.76)	23.48 (3.93)	1.238	.223
Empathy (SD)	36.13 (8.03)	37.45 (9.85)	35.62 (7.36)	-.640	.526
Between Group Comparison T2 – 2nd Assessment					
	Total score	Pathological gamers (n=11)	Healthy controls (n=29)	t-value	p-value
Age (SD)	22.34 (2.12)	22.72 (2.94)	22.19 (2.37)	-.597	.554
AICA_30 (SD)	2.65 (2.94)	4.63 (3.59)	1,90 (2,30)	-2.865	.007

AICA_lifetime (SD)	13.65 (9.33)	23.19 (6.19)	10.03 (7.64)	-5,094	<.001
Years of education (SD)	-	-	-	-	-
SASKO: Speaking (SD)	7.78 (5.04)	9.63 (5.85)	6.86 (4.56)	-1.589	.120
SASKO: Rejection (SD)	6.81 (4.71)	9.45 (6.15)	6.04 (3.71)	-2.152	.038
SASKO: Interaction (SD)	5.50 (3.58)	5.91 (3.78)	5.41 (3.50)	-.377	.711
SASKO: Information (SD)	6.18 (3.91)	7.82 (4.40)	5.66 (3.60)	-1.456	.119
SASKO: Loneliness (SD)	1.84 (2.42)	2.36 (2.11)	1.72 (2.47)	-.814	.454
EKF-EE (SD)	58.05 (9.21)	57.36 (11.78)	58.33 (8.19)	.291	.773
EKF-EA (SD)	62.06 (14.27)	58.32 (22.32)	63.59 (7.21)	.757	.464
EKF-RE (SD)	51.08 (6.94)	53.09 (7.57)	50.26 (6.64)	-1.146	.259
EKF-EX (SD)	51.98 (13.92)	51.12 (20.29)	52.33 (10.88)	.240	.812
Rosenberg Selfworth (SD)	23.76 (4.76)	23.09 (3.70)	24.04 (4.01)	.551	.585
Empathy (SD)	36.14 (9.75)	39.10 (9.99)	35.00 (9.61)	.803	.265
Within-group comparisons T1 > T2					
	Pathological gamers t-value	p-value	Healthy con-trols t-value	p-value	
Age (SD)	-9.037	<.001	-13.25	<.001	
AICA_30 (SD)	4.087	.002	.462	.647	
AICA_lifetime (SD)	-.396	.700	.405	.688	
Years of education (SD)	-	-	-	-	
SASKO: Speaking (SD)	.840	.421	1.204	.238	
SASKO: Rejection (SD)	.897	.391	1.528	.138	
SASKO: Interaction (SD)	1.168	.270	.328	.745	
SASKO: Information (SD)	.482	.640	-.982	.334	
SASKO: Loneliness (SD)	.510	.563	1.063	.297	
EKF-EE (SD)	-.398	.699	-.400	.692	
EKF-EA (SD)	.467	.651	.029	.977	
EKF-RE (SD)	-2.444	**.035**	-1.090	.285	

EKF-EX (SD)	-.100	.922	-.410	.685
Rosenberg Selfworth (SD)	-2.000	.074	-.583	.565
Empathy (SD)	-.425	.681	.782	.442

AICA_30: Severity of computer game addiction during the last 30 days, AICA_Lifetime: lifetime usage of computer games, SD: standard deviation, SASKO: Social Anxiety and Social Competence Deficits, EKF-EE: Recognizing and understating own emotions, EKF-EA: Recognizing and understanding others' emotions, EKF-RE: Regulation and control of own emotions, EKF-EX: Emotional expressiveness; t= two sample t-test statistics

Table 2.7 Intraclass correlation coefficients of self-concept-related measures

	Total Sample N=40		**Pathological gamers (n=11)**		**Healthy controls (n=29)**	
	Intraclass Correlatic coefficien	p-value	Intraclass Correlation coefficient	p-value	Intraclass Correlation coefficient	p-value
SASKO: Speaking	.820	<.001	.810	.001	.809	<.001
SASKO: Rejection	.762	<.001	.843	<.001	.665	<.001
SASKO: Interaction	.631	<.001	.730	.003	.545	.001
SASKO: Information	.727	<.001	.668	.009	.728	<.001
SASKO: Loneliness	.845	<.001	.664	.009	.908	<.001
EKF-EE	.552	<.001	.881	<.001	.390	.017
EKF-EA	.481	.001	.319	.156	.717	<.001
EKF-RE	.461	.001	.478	.058	.502	.002
EKF-EX	.709	<.001	.712	.005	.705	<.001
Rosenberg Selfworth	.787	<.001	.905	<.001	.679	<.001
Empathy	.464	.002	.475	.070	.453	.009

SASKO: Social Anxiety and Social Competence Deficits, EKF-EE: Recognizing and understating own emotions, EKF-EA: Recognizing and understanding others' emotions, EKF-RE: Regulation and control of own emotions, EKF-EX: Emotional expressiveness

Group-level brain activation

Analyses of brain activation across both groups (N=40) indicated significant self-concept associated brain activation (contrast: "self > familiar and unknown person") in the bilateral insula; the anterior and medial cingulum; the IFG; the operculum; the bilateral putamen; the claustrum; the superior motor area; the precentral gyrus and the STG; the right globus pallidus; the superior and medial frontal gyri; the supramarginal gyrus; the postcentral gyrus; and the inferior parietal lobe (see Supplementary Table S 2.6). The patterns of brain activation were replicated, when analyzing both groups separately (see Supplementary Table S 2.6). Between-group comparisons did not reveal significant differences in brain activations at T1 or T2. Longitudinal comparison of the brain activation during first and second fMRI did not show a significant change in brain activation over time (two-tailed: increase over time [T1<T2] or decrease [T1>T2]; see Supplementary Table S 2.6). Detailed results of the group-level analyses for the constituting task conditions (self, familiar person, and unknown person) are depicted in the Supplementary Figure S 2.7. In short, self-related activation patterns were detected in both groups in the insula as well as in the superior and inferior temporal gyrus. For the familiar and unknown condition, activation was recorded in the middle and superior temporal gyrus.

Reliability analyses

Reliability of the contrast "self > familiar and unknown person"

For the pooled study sample (N=40) and in both study groups separately, the mean ICC values of the contrast "self > familiar and unknown person" were under the threshold of moderate reliability (ICC<0.4, see Table 2.8). This finding, however, came as no surprise as we had assumed that brain activation in areas, which are unrelated to the fMRI task, and the construct of self-evaluation could not be replicated in its magnitude. This resulted in a low overall ICC value across the whole brain. Both groups did not significantly differ in mean ICC values. Thresholded ICC maps illustrated that the local reliability of the contrast "self > familiar and unknown person" did not surpass the threshold for good reliability, neither in the pooled sample, nor when considering

both experimental groups separately (ICC>0.75, see Figure 2.7 and Supplementary Figure S 2.8 and Supplementary Figure S 2.9; ICC maps are available at https://identifiers.org/neurovault.collection:9777). However, several brain areas showed moderate to good reliability (0.75>ICC>0.60, see Figure 2.8). In the patient group these areas included the right middle and anterior cingulum; the right superior temporal gyrus, including parts of the TPJ; the bilateral middle and inferior temporal gyrus; the left fusiform gyrus; the left insula; and the right inferior occipital gyrus. In the control group the areas included the bilateral middle occipital gyri as well as the right superior and middle temporal gyrus, including parts of the TPJ.

The atlas-based summary of mean ICC values for 120 brain regions for the contrast "self > familiar and unknown person" (collapsed across both groups) showed that no brain region surpassed an average mean ICC value of 0.4 (see Supplementary Table S 2.7).

The overall within-subject similarity for all contrast conditions exceeded the between subject similarity values (within: $r_{self} = 0.50$, $r_{familiar\ other} = 0.19$, $r_{unkown\ other} = 0.52$, $r_{self\text{-}other} = 0.23$; between: $r_{self} = 0.28$, $r_{familiar\ other} = 0.04$, $r_{unkown\ other} = 0.32$, $r_{self\text{-}other} = 0.09$). This translated into a high proportion of participants that could be re-identified based on their neural signature during the different task conditions (i.e. participants can be re-identified if their within-subject correlation coefficients exceeded all between-subject correlation coefficients with other participants). 55% of participants could be re-identified based on their activation during the "self" condition and 69%, when investigating the "unknown other" contrast, while 36% of the sample could be re-identified based on the activation during the "familiar other" condition, and still 30% of participants could be re-identified using the difference contrast ("self-other").

Results of the similarity analysis demonstrate higher within-subject similarity compared to between-subject similarity from first to second fMRI for the contrast "self > familiar and unknown person" ($t_{self\text{-}familiar+unknown\ other}>=5.37$, $t_{self}>=9.93$, $p<0.001$). In both groups, the overall within-subject similarity exceeded the mean between-subject similarity values (within: $r_{self\text{-}\ familiar+unknown\ person\ r}>=0.16$; between: $r_{self\text{-}\ familiar+unknown\ person}$

<=0.10). This translated into a proportion of participants that could be re-identified based on their neural signature captured by the contrast "self > familiar and unknown person" of 30% (pooled sample, N=40) with a higher proportion in the group of healthy participants (37%), compared to the groups of individuals with problematic internet use (27%, see Figure 2.9 and Figure 2.10).

The low overall reliability of the "self > familiar and unknown person" contrast also reflected in low Jaccard and Dice coefficients, in the whole study sample and also when analyzing both study groups separately (see Table 2.9), which indicated that only about 1–8% of significant voxels could be replicated during the second fMRI session. Analyses showed that the magnitude of Jaccard and Dice reliability estimates was significantly lower for the "self > familiar and unknown person" contrast, compared to the three constituent task conditions ($F_{Jaccard(3,114)}$=54.386, p<0.001, eta^2=0.589; $F_{Dice(3,114)}$=64.886, p<0.001, eta^2=0.631). There was no significant difference between pathological gamers and controls ($F_{(1,38)Jaccard}$=0.453, p=0.505, eta^2=0.012; $F_{(1,38)Dice}$=0.355, p=0.555, eta^2=0.009).

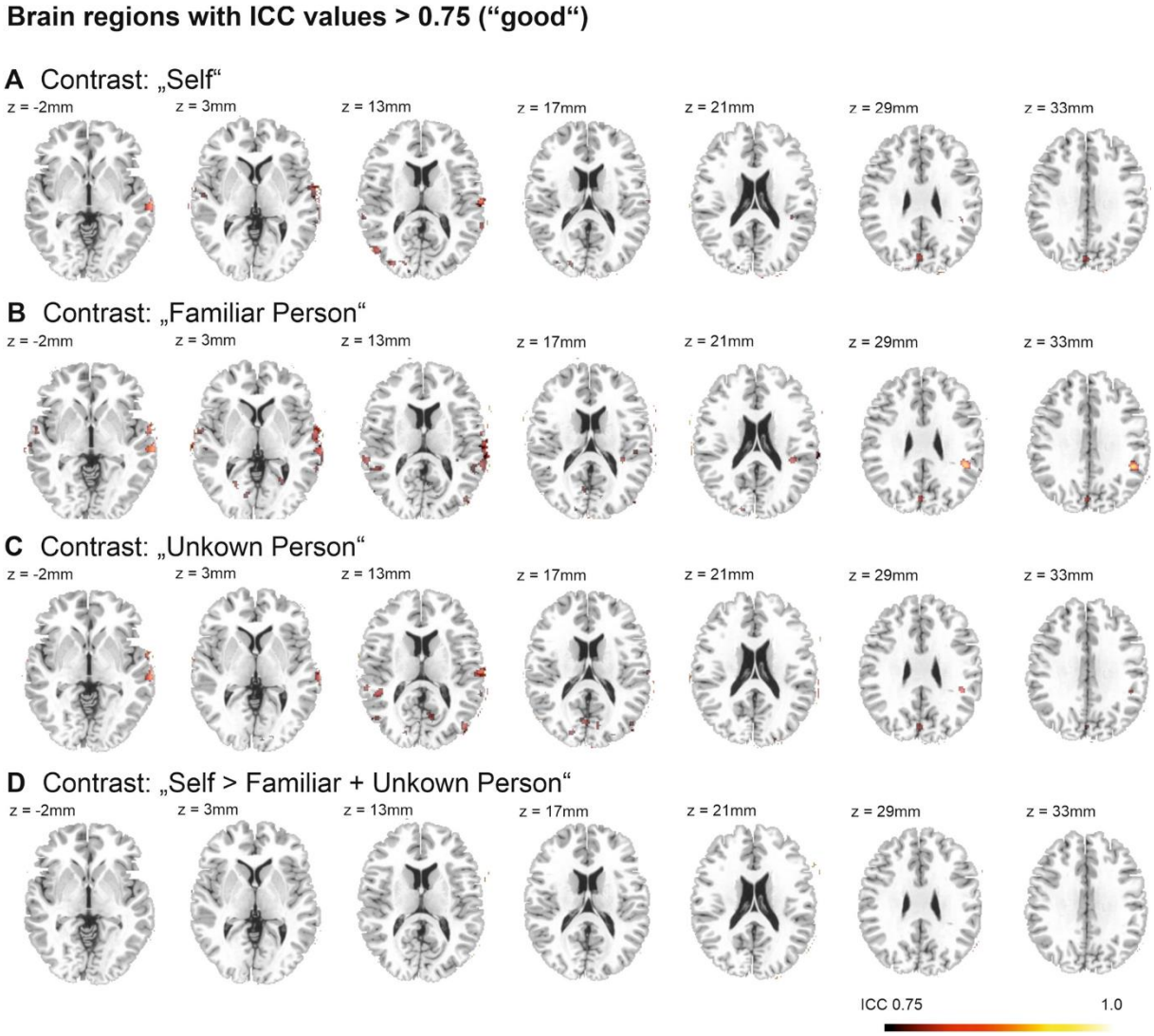


Figure 2.7 Depiction of brain areas that show good to excellent reliability for the different task contrasts: A] "Self", B] "Familiar Person", C] "Unknown Person" and D] "Self > Familiar + Unknown Person" (Intraclass correlation coefficient [ICC] > 0.75), when performing pooled analyses of the whole dataset of N=40 participants.

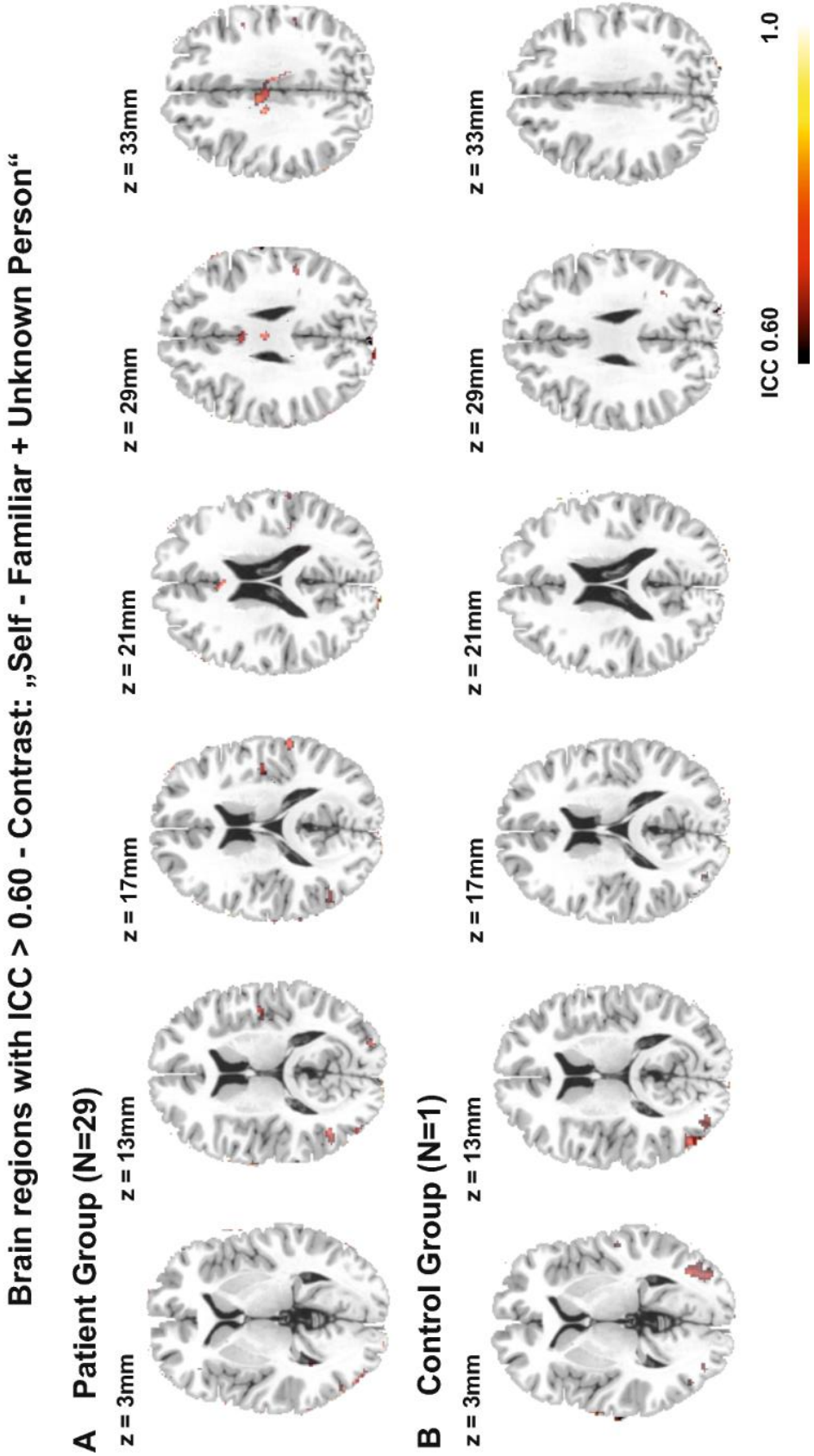


Figure 2.8 Depiction of brain areas that show moderate to good reliability (0.75>ICC>0.60) for the contrast "self > familiar and unknown person" in (A) the patient group and (B) the control group.

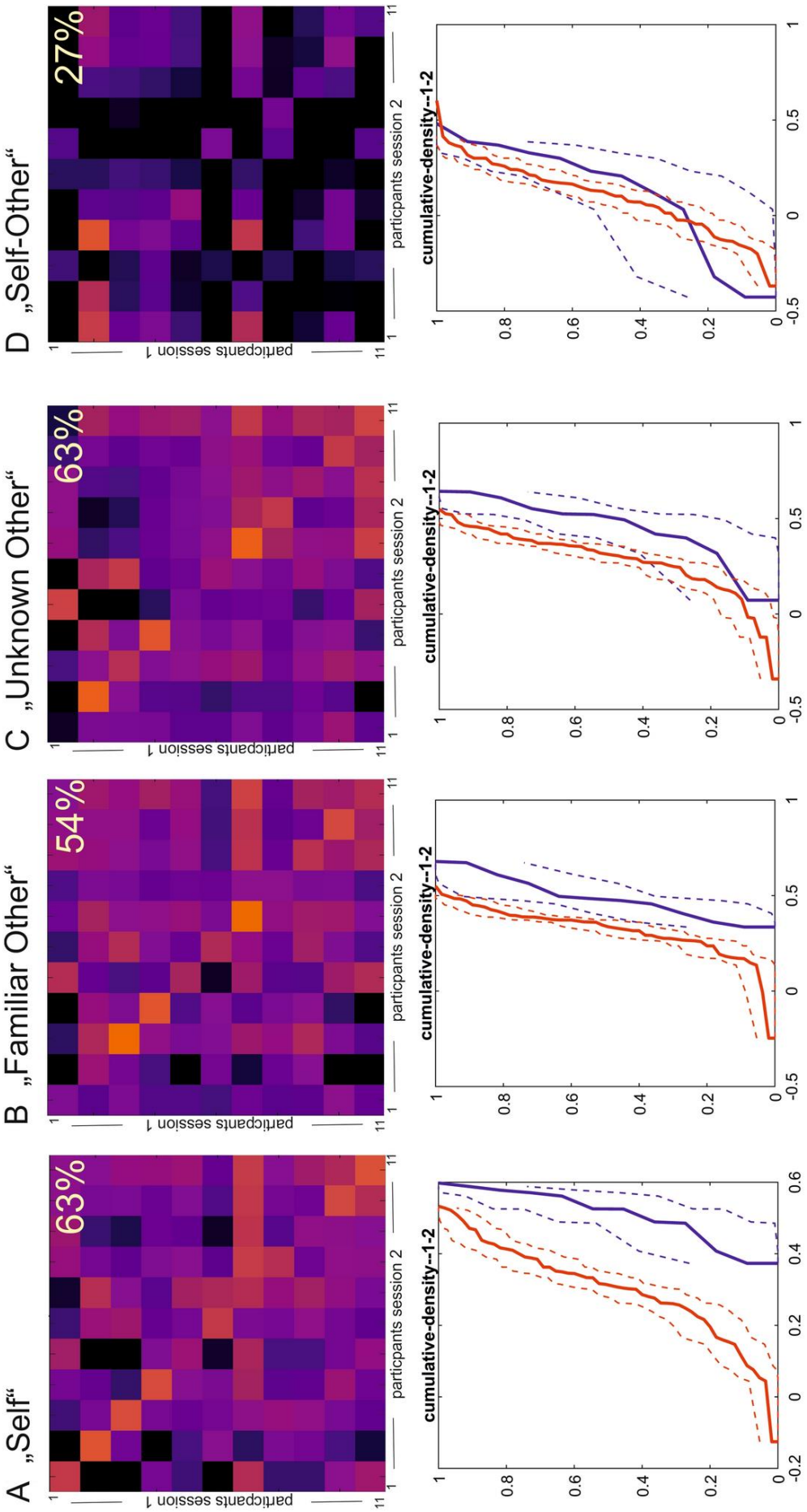

A „Self“
63%
B „Familiar Other“
54%
C „Unknown Other“
63%
D „Self-Other“
27%
participants session 1
participants session 2
cumulative-density--1-2

Figure 2.9 Similarity maps for the problematic gamers group (upper row) and empirical cumulative distribution functions (lower row, red lines: between-subject similarity: lower row, blue lines: within-subject similarity) for longitudinal comparisons (first and second fMRI sessions) for the four contrast conditions: A] "self", B] "familiar person", C] "unknown person", and D] "self > familiar and unknown person". The diagonal of each color matrix represents the within-subject similarity values. Re-identification of a subject based on the neural activation map is affirmed 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 values (red lines) for the A] "self", B] "familiar person", and C] "unknown person" contrast maps; hereby, the cumulative density functions overlapped for D] the "self > familiar and unknown person" contrast.

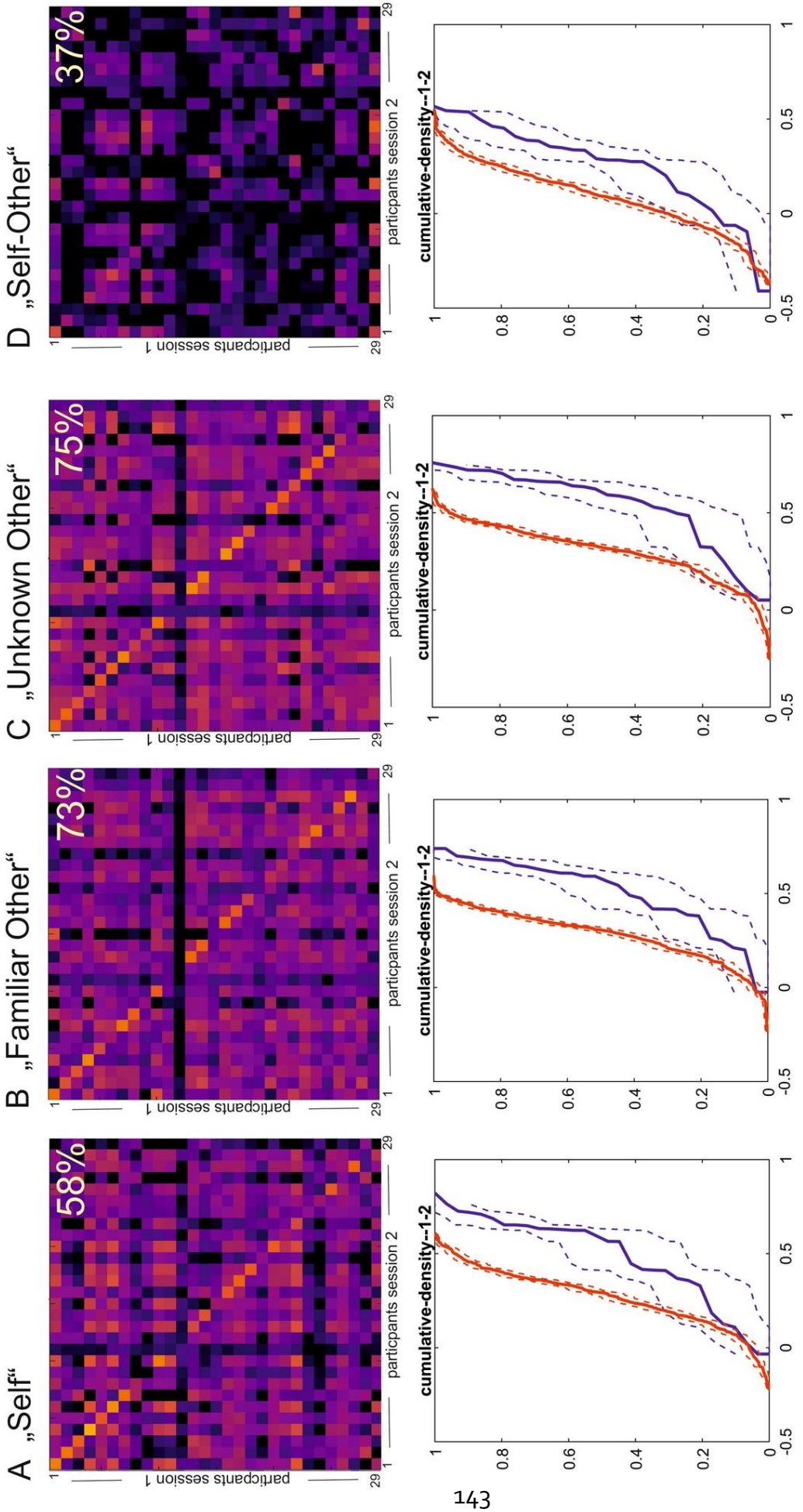

A „Self“
58%
B „Familiar Other“
73%
C „Unknown Other“
75%
D „Self-Other“
37%
participants session 1
participants session 2
cumulative-density--1-2

Figure 2.10 Similarity maps for the control group (upper row) and empirical cumulative distribution functions (lower row, red lines: between-subject similarity; lower row, blue lines: within-subject similarity) for longitudinal comparisons (first and second fMRI sessions) for the four contrast conditions: A] "self", B] "familiar person", C] "unknown person", and D] "self > familiar and unknown person". The diagonal of each color matrix represents the within-subject similarity values.

Table 2.8 Mean Intraclass Correlation (ICC) values for different contrast conditions for the pooled sample and both study groups separately.

		Contrasts			
		Self	**Familiar Person**	**Unknown Person**	**Self > Familiar + Unknown Person**
Pooled group (N=40)	**Mean ICC (SD)**	0.24 (.24)	0.30 (.24)	0.25 (.24)	0.06 (.21)
Controls (N=29)	**Mean ICC (SD)**	0.25 (.24)	0.28 (.24)	0.27 (.24)	0.07 (.21)
Patients (N=11)	**Mean ICC (SD)**	0.25 (0.23)	0.24 (0.22)	0.21 (0.21)	-0.03 (0.18)

Table 2.9 Comparison of Jaccard and Dice coefficients across the different task conditions for the pooled sample and both study groups separately.

			Contrasts					
			Self	**Familiar Person**	**Unknown Person**	**Self > Familiar + Unknown Person**	**Statistics**	***p***
Pooled group (N=40)	**Jaccard Coefficient**	**Mean**	0.24	0.30	0.32	0.04	$F_{(3,84)}$=78.766	<0.001
		(SD)	(0.16)	(0.14)	(0.15)	(0.09)		
	Dice Coefficient	**Mean**	0.36	0.44	0.47	0.07	$F_{(3,84)}$=89.051	<0.001
		(SD)	(0.22)	(0.18)	(0.20)	(0.14)		
Controls (N=29)	**Jaccard Coefficient**	**Mean**	**0.25**	**0.31**	**0.34**	**0.05**	$F_{(3,84)}$=41.825	<0.001
		(SD)	(0.16)	(0.14)	(0.15)	(0.09)		
	Dice Coefficient	**Mean**	**0.38**	**0.45**	**0.49**	**0.08**	$F_{(3,84)}$=48.060	<0.001
		(SD)	(0.21)	(0.19)	(0.19)	(0.14)		
Patients (N=11)	**Jaccard Coefficient**	**Mean**	**0.37**	**0.43**	**0.47**	**0.02**	$F_{(3,30)}$=29.128	<0.001
		(SD)	(0.23)	(0.12)	(0.15)	(0.03)		
	Dice Coefficient	**Mean**	**0.24**	**0.28**	**0.32**	**0.01**	$F_{(3,30)}$=34.137	<0.001
		(SD)	(0.17)	(0.09)	(0.12)	(0.02)		

Post-hoc tests demonstrated a significant difference between the "self > familiar and unknown person" contrast and all other contrast conditions (all p's<0.003), while the (i) self, (ii) familiar person, and (iii) unknown person contrast conditions did not differ in the magnitude of the Dice and Jaccard coefficients (all p's >0.05).

Reliability of the constituent task conditions "self", "familiar person", and "unknown person"

The mean ICC values for the (i) self, (ii) familiar person, and (iii) unknown person contrasts were remarkably higher, compared to the contrast "self > familiar and unknown person" in the pooled sample and also when analyzing both study groups separately (see Table 3). Still, the mean ICC values for the pooled dataset (N=40) were below the threshold for moderate reliability (0.4). Thresholded ICC maps for the (i) self, (ii) familiar person, and (iii) unknown person contrast demonstrated good to excellent reliability (1.0>ICC>0.75, see Figure 2.7) in several brain areas. In the patient group these areas included parts of the bilateral superior; middle and inferior temporal gyrus, including parts of the TPJ; the left and right fusiform gyrus; the right angular gyrus; the middle and inferior occipital gyrus; precuneus; cuneus; superior frontal gyrus; postcentral and precentral gyrus; amygdala; pallidum; insula; cerebellum and parts of the left lingual gyrus; the fusiform gyrus; middle and inferior occipital gyrus; superior, middle, and inferior frontal gyrus; precentral and postcentral gyrus; cuneus; calcarine; and caudate (see supplementary Figure S2). In the control group these areas included parts of the right middle and inferior occipital gyrus; the superior and middle temporal gyrus; cuneus; lingual gyrus; calcarine; precuneus and the left superior; middle and inferior temporal gyrus; the postcentral gyrus; precentral gyrus; Heschl gyrus; supramarginal gyrus; insula; the superior and inferior frontal gyrus and superior occipital gyrus (see Supplementary Figure S 2.9).

The atlas-based summary of mean ICC values for the contrast conditions (i) self, (ii) familiar person, and (iii) unknown person showed that several brain regions surpassed the threshold of fair to moderate reliability (ICC>0.4). These included the bilateral superior; middle and inferior occipital gyri; the superior and middle temporal gyri; the Heschl gyri; the supramarginal gyri; as well as the cuneus, calcarine and lingual gyri (see Supplementary Table S 2.7).

Results of the three constituting contrast conditions show higher within-subject similarity than between-subject similarity from the first to the second fMRI. The difference

between within-subject and between-subject similarity is depicted as the prominent diagonal in the similarity matrices (see Figure 2.9and Figure 2.10). In both groups, the overall within-subject similarity for all contrast conditions exceeded the between-subject similarity values (within: $r_{self} >= 0.50$, $r_{familiar\ person} = 0.50$, $r_{unknown\ person} = 0.47$; between: $r_{self} <= 0.29$, $r_{familiar\ person} <= 0.31$, $r_{unknown\ person} <= 0.33$). Analyses based on the pooled dataset (N=40) indicated that 55% of participants could be re-identified based on their activation during the "self" contrast, 69% could be re-identified based on their activation during the "unknown person" contrast and 36% could be re-identified based on their activation during the "familiar person" contrast. Considering the two experimental groups separately, these results showed comparable results. 63% of the participants in the patient group and 58% of the participants in the control group could be re-identified based on their activation during the "self" contrast condition. 54% and 72% respectively could be re-identified based on the "unknown person" contrast. 63% and 75% respectively could be re-identified based on the activation during the "familiar person" condition (see Figure 2.9 and Figure 2.10).

Analyses of the Jaccard and Dice coefficients for the pooled dataset and considering both experimental groups separately demonstrated a significant main effect in the contrast category for both coefficients, while there was no significant difference between groups. The "self", "familiar person", and "unknown person" contrast conditions displayed significantly higher Dice and Jaccard coefficients compared to the "self > familiar and unknown person" contrast (all p's <=0.003, see Table 2.9). About 24–47% of the significant voxels for "self", "familiar person", and "unknown person" contrasts were replicated during the second fMRI session.

Assessment of factors underlying the reliability differences across the contrast conditions

In order to assess whether the comparatively low reliability of the contrast ("self > familiar and unknown person") might result from an inter-correlation between the constituting single task condition contrasts, we computed the voxel-wise Pearson correlation coefficient for all three constituting conditions. Data indicate substantial correlations between the three conditions ($r_{self\ x\ familiar\ person} = 0.36$, SD = 0.26, $R^2=0.13$, $r_{self\ x}$

$r_{unknown\ person}$ = 0.28, SD = 0.32, R^2=0.08, $r_{unknown\ person\ x\ familiar\ person}$ = 0.29, SD = 0.35, R^2=0.08). This evinces that the three constituting task conditions share about 8–12% of their variance. Previous studies indicate that part of the shared variance is removed by subtracting the constituting contrast conditions. In our case, this is illustrated by the lower correlation coefficients between the "self vs. familiar and unknown person" contrast and the constituting contrast conditions ($r_{(self-\ familiar+unknown\ person)\ x\ self}$ =0.22, SD=0.30, R^2=0.05; $r_{(self-\ familiar+unknown\ person)\ x\ familiar\ person}$=-0.01, SD=0.18, R^2<0.01; $r_{(self-\ familiar+unknown\ person)\ x\ unknown\ person}$ =-0.16, SD=0.24, R^2=0.03).

2.3.5 Discussion

Our study assessed whole-brain longitudinal reliability of an fMRI task that was designed to investigate the neurobiological correlates of participants' self-concept.

First, we assessed the robustness of group-level brain activation for the task contrast of interest ("self vs. familiar and unknown person"). Results indicate stable brain activation in the bilateral insula; the anterior and medial cingulate; as well as the IFG in both groups at T1. This is in line with the meta-analytical findings of Hu et al. (2016) who report that the ACC, the insula, the IFG as well as regions of the TPJ are activated during reflections and recognition of the own person as compared to other individuals. Furthermore, analyses of psychometric measures that assessed aspects of the self-concept indicated that the vast majority of measures did not show a significant change over time, with moderate to high test-retest reliability of the measures. This supports the notion that self-concept-related aspects can be regarded as relatively stable, thus, confirming the stability of the self-concept construct (Rammsayer & Weber, 2016; Swarm Jr, 1983).

Reliability analyses for the contrast of interest "self > familiar and unknown person" showed poor overall reliability, indicated by low mean ICC values and low Jaccard and Dice coefficients. Still, local ICC values indicated fair to moderate reliability in parts of the bilateral middle occipital gyri; the middle and superior temporal gyri; and parts of the TPJ. Similarity analysis indicated that a quarter to a third of the participants could be re-identified based on their neural activation pattern encoded by the contrast "self

> familiar and unknown person". This value might seem unexpectedly high considering the low ICC, Dice, and Jaccard coefficients. However, the very low between-subject correlation values for the contrast in combination with the moderate within-subject correlation coefficients resulted in a re-identification of a substantial proportion of the sample (i.e., the within-subject correlation is higher than all between-subject correlation coefficients). In order to determine whether the results depended on the sample size, we conducted reliability analyses with the pooled datasets of n=40 participants, in addition to the analyses for the two experimental groups separately. Results of the pooled dataset (N=40) confirm the findings derived from the separate group analyses.

In opposition to the low overall reliability of the "self > familiar and unknown person" contrast, the single constituting contrast conditions "self", "familiar person", and "unknown person" showed good to excellent longitudinal reliability in several brain regions associated with the processing of self-referential information (Goldin et al., 2014; Hu et al., 2016). Specifically, this included the bilateral superior and middle temporal and occipital gyri; portions of the TPJ; as well as parts of the cuneus, lingual gyrus, calcarine and areas of the mesolimbic system, such as the insula. The Dice and Jaccard coefficients indicated that about a quarter to a third of the clusters displaying significant activation for the "self", "familiar person", and "unknown person" contrasts during the baseline fMRI assessment could be replicated in the second fMRI assessment after one year. Similarity analyses also showed that more than half of the participants could be re-identified based on neural activation patterns.

The lower reliability of the contrast "self > familiar and unknown person" might result—at least in part—from a substantial correlation between the constituting contrast conditions. These conditions share about a tenth of their variance, as indicated by the Pearson correlation coefficients. This proportion of shared variance is removed by subtracting the conditions from one another, while the error variances are added. In their recent publication, Infantolino et al. (2018) confirmed that the correlation between the constituting contrast conditions of a contrast, places an upper limit on the

reliability of a difference measure. Summarizing the results of 56 independent fMRI studies, a recent meta-analysis showed that only half of the reliability scores fell within the range of at least moderate reliability (Elliott et al., 2020). The authors concluded that difference scores will always have lower reliability than their constituent contrast conditions; hence, limiting the reliability of such a measure (Elliott et al., 2020).

The sub-perfect reliability of the constituting task condition contrasts points out that additional factors underlie the observed limited reliability. Elliot et al. (2020) investigated factors that might limit reliability of task fMRI. Their moderator analysis indicated that neither task type (i.e., process under investigation); task design (e.g., block vs. event-related); task length; test-retest interval; ROI type (i.e., structural vs. functional); nor sample type (i.e., healthy vs. clinical) significantly moderated reliability scores. In the presented analyses, we could not determine the individual factors underlying the sub-perfect reliability of the constituting task condition contrasts. However, it is likely that the self-concept—as well as the associated cognitive processes that are captured by the task under investigation—vary over the test-retest interval of one year, limiting the overall reliability. Still, we observed good local reliability of the constituent task conditions, which suggests that fMRI-based measures of the presented task can provide sufficiently reliable estimates of brain activation. In regard to the poor reliability of the contrast "self > familiar and unknown person", several steps can be undertaken to mitigate the problems associated with computing the contrast between the constituent task conditions. In the case of a linear association and correlation between the constituent task conditions, one task condition can substitute the other condition without losing information on the individual differences between participants (Infantolino et al., 2018). Regarding the self-evaluation task, we argue that the individual response trajectories to pictures of oneself are of special interest. We assume that changes in altered self-concept translates into changes in brain activation captured by the "self" contrast. It can be argued that the focus on a solitary task condition reduces the capacity to isolate specific cognitive processes. This could be overcome by either relying on meta-analysis that could inform on the role of a certain brain region for the cognitive process under investigation. Alternatively, the investigation

of within-subject effects between the constituting task conditions could be used to identify regions that specifically activate differently to various task conditions and regions of interest. In a next step, brain activation during the constituting condition "self" could be used as a measure to index individual differences over time with moderate to good or even excellent reliability.

Limitations

The small sample size of the group of participants with problematic internet gaming use and IGD should be considered a relevant limitation. In addition, the meta-analysis by Elliott et al. (2020) reported that the sample size of studies assessing fMRI reliability ranged from five to 58 subjects with a median below 30. While the sample size of the current study should be regarded as a potential limitation, the sample size does, in fact, exceed most previous fMRI reliability studies. Still, the presented findings should be regarded as preliminary and future studies with larger sample sizes are needed.

2.3.6 Conclusion

To the best of our knowledge, this is the first study that presents longitudinal reliability analyses of a self-evaluation fMRI paradigm. Self-concept-associated brain activation, indexed by the contrast "self > familiar and unknown person" showed poor overall reliability. Still, local reliability measures demonstrated good reliability in regions of the TPJ. Furthermore, similarity analyses indicated that about a third of the participants could be re-identified based on their neural activation, captured by the contrast "self > familiar and unknown person". In contrast, the reliability estimates consistently indicated good to excellent reliability of the constituting task conditions "self", "unknown person", and "familiar person" in several brain regions associated with social cognition. The poor global reliability of the contrast "self > familiar and unknown person" could be explained—at least in part—with a substantial correlation between the brain activation of the constituting contrast conditions. Future fMRI research on self-evaluation should be cautioned by these findings and employ methods to overcome these limitations.

2.3.7 Supplements

Supplementary Tables

Supplementary Table S 2.6 Brain areas depicting significantly higher activating during viewing videos of oneself compared to videos of other persons (contrast: "self > familiar + unknown person", whole-brain threshold p<.001, pFWE, Cluster < .05).

					MNI coordinates			
H	***Lobe***	***BA***	***Brain regions***	***Cluster size***	***X***	***Y***	***Z***	***T***
Pathological gamers (N=11) T1								
L	Frontal	13/44	Insula, inferior frontal Gyrus, Gyrus Precentralis	577	-40	8	4	5.89
R/L	Limbic/Frontal	32/24	Anterior and medial Cingulum	530	-2	38	8	4.75
R	Frontal	13/47	Insula, inferior frontal Operculum, Putamen, inferior Frontal Gyrus	611	36	16	10	4.45
Controls (N=29) T1								
L	Frontal/ Temporal	13/22	Insula, inferior frontal Gyrus, inferior frontal Triangularis, superior temporal gyrus, Claustrum, Gyrus precentralis	1149	-42	12	-2	6.66
R	Frontal	45/47	Insula, inferior frontal Gyrus, inferior frontal Operculum, Gyrus Precentralis, Putamen, Claustrum	834	30	24	6	6.26

L/R	Limbic/ Frontal	32/24	Anterior and medial Cingulum	1129	4	28	26	4.93
All (N=40) T1								
L	Frontal	13/22/38/44/45/47	Insula, inferior frontal Gyrus Precentralis, Putamen, superior temporal Gyrus, Claustrum, inferior frontal Operculum,	1651	-40	10	2	8.1
R	Frontal/temporal	13/22/44/45/47	Insula, inferior frontal Gyrus, inferior frontal Operculum, Putamen, Gyrus Precentralis, Globus Pallidum, Superior Temporal Gyrus, Claustrum	1764	32	20	8	6.43
L/R	Limbic/Frontal	24/9/32	Anterior and medial Cingulum (r and l), superior and medial frontal Gyrus (r), superior Motor Area (r and l)	1807	2	28	26	6.36
L/R	Frontal	6/8	Superior and medial frontal Gyrus (r), superior Motor Area (r and l)	585	-4	6	62	5.99
R	Parietal	1/2/3/40	Gyrus supramarginalis, Gyrus postcentralis, inferior parietal Lobe, Gyrus Precentralis	345	56	-30	38	5.35

Pathological gamers (N=11) T2
n.s.
Controls (N=29) T2
n.s.
All (N=40) T2
n.s.

Pathological gamers (N=11) T2>T1 and T2<T1
n.s.
Controls (N=29) T2>T1 and T2<T1
n.s.
All (N=40) T2>T1 and T2<T1
n.s.

Note. H = hemisphere; L = left; R = right; BA = Brodmann area; MNI = Montreal Neurological Institute

n.s. = not significant; Whole-Brain Threshold $p<.001$, $p_{FWE,\ Cluster} < .05$).

Supplementary Table S 2.7 Atlas-based mean Intraclass Correlation (ICC) values for the four task contrasts "self", "familiar other", unknown other" and "self – other" for 120 anatomical regions specified in the aal atlas for the pooled analyses of the whole sample (N=40).

	Contrasts			
Brain Region (aal atlas)	**Self**	**Familiar Person**	**Unknown Person**	**Self > Familiar + Unknown Person**
2001_Precentral_L	0.27	0.38	0.37	0.03
2002_Precentral_R	0.37	0.35	0.33	0.00
2101_Frontal_Sup_2_L	0.22	0.28	0.23	0.05
2102_Frontal_Sup_2_R	0.25	0.29	0.28	-0.01
2201_Frontal_Mid_2_L	0.17	0.32	0.21	-0.01
2202_Frontal_Mid_2_R	0.20	0.29	0.26	-0.02
2301_Frontal_Inf_Oper_L	0.23	0.39	0.28	0.09
2302_Frontal_Inf_Oper_R	0.19	0.27	0.21	-0.03
2311_Frontal_Inf_Tri_L	0.34	0.37	0.23	0.09
2312_Frontal_Inf_Tri_R	0.22	0.32	0.14	0.00
2321_Frontal_Inf_Orb_2_L	0.21	0.28	0.17	0.05
2322_Frontal_Inf_Orb_2_R	0.08	0.29	0.17	-0.06
2331_Rolandic_Oper_L	0.40	**0.53**	**0.40**	0.09
2332_Rolandic_Oper_R	0.37	**0.46**	0.29	0.03
2401_Supp_Motor_Area_L	0.36	0.35	0.30	0.01
2402_Supp_Motor_Area_R	0.32	0.37	0.29	-0.01
2501_Olfactory_L	0.23	0.19	0.23	-0.06
2502_Olfactory_R	0.23	0.19	0.21	-0.13
2601_Frontal_Sup_Medial_L	0.21	0.30	0.24	0.04
2602_Frontal_Sup_Medial_R	0.16	0.26	0.25	-0.01

2611_Frontal_Med_Orb_L	0.07	0.22	0.09	0.00
2612_Frontal_Med_Orb_R	0.03	0.14	0.06	-0.06
2701_Rectus_L	-0.04	0.26	0.11	-0.07
2702_Rectus_R	0.02	0.18	0.07	-0.18
2801_OFCmed_L	-0.01	0.24	0.07	-0.15
2802_OFCmed_R	0.04	0.21	0.09	-0.06
2811_OFCant_L	0.00	0.25	0.13	-0.07
2812_OFCant_R	-0.01	0.20	0.09	-0.02
2821_OFCpost_L	0.02	0.21	0.06	-0.10
2822_OFCpost_R	0.07	0.27	0.15	-0.09
2831_OFClat_L	0.03	0.32	0.12	-0.02
2832_OFClat_R	0.08	0.26	0.02	0.02
3001_Insula_L	0.10	0.28	0.17	-0.02
3002_Insula_R	0.12	0.29	0.16	-0.04
4001_Cingulate_Ant_L	0.08	0.25	0.18	0.03
4002_Cingulate_Ant_R	-0.02	0.24	0.16	0.06
4011_Cingulate_Mid_L	0.15	0.36	0.21	-0.02
4012_Cingulate_Mid_R	0.18	0.33	0.24	0.01
4021_Cingulate_Post_L	0.20	0.36	0.19	0.17
4022_Cingulate_Post_R	0.09	0.30	0.24	0.16
4101_Hippocampus_L	0.07	0.24	-0.02	0.07
4102_Hippocampus_R	0.06	0.19	-0.01	-0.01
4111_ParaHippocampal_L	0.10	0.22	0.14	-0.02
4112_ParaHippocampal_R	0.09	0.23	0.11	-0.06
4201_Amygdala_L	0.20	0.32	0.17	0.04
4202_Amygdala_R	0.16	0.28	0.08	-0.08
5001_Calcarine_L	**0.47**	**0.57**	**0.56**	0.18

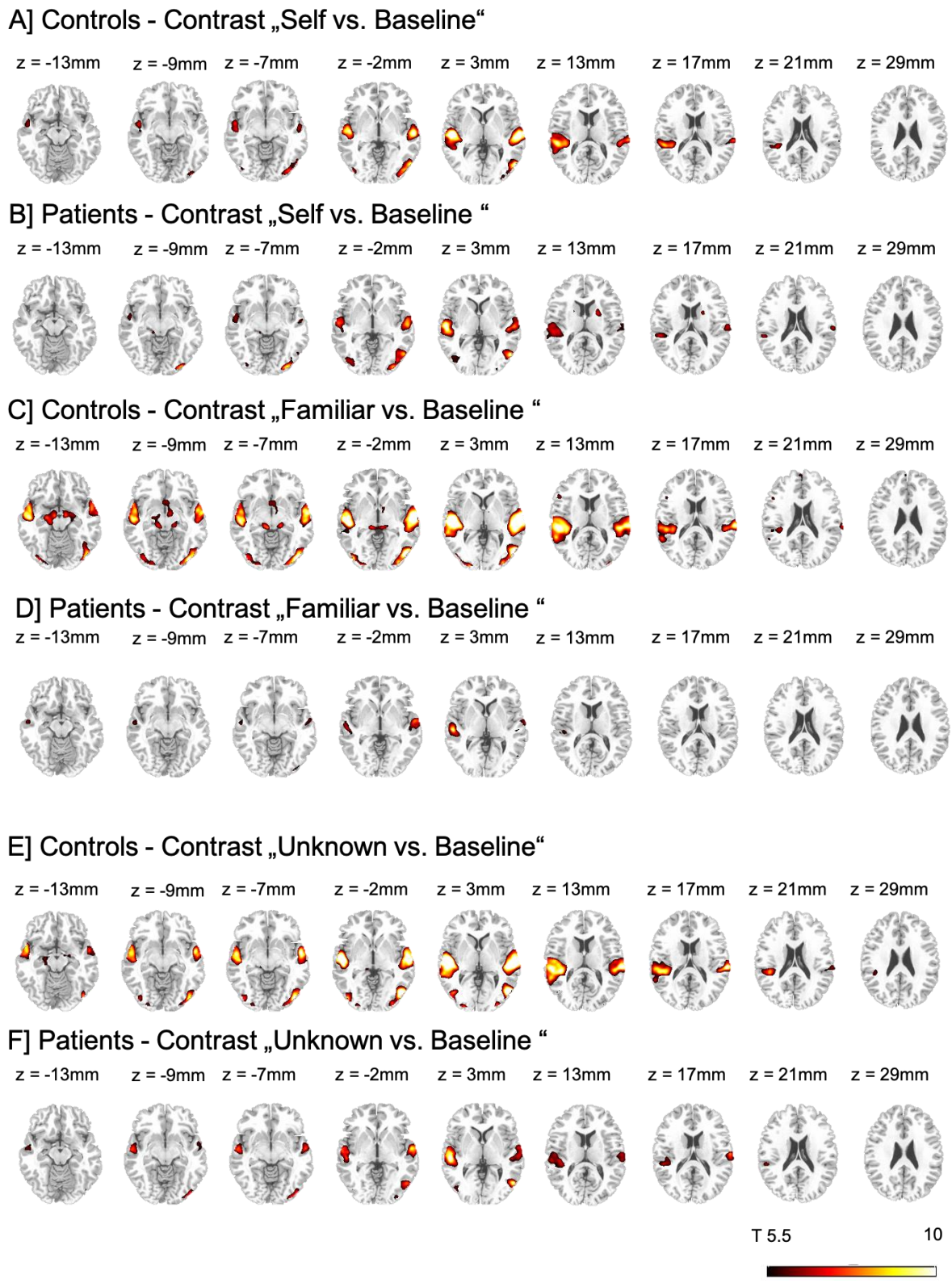
5002_Calcarine_R	**0.54**	**0.61**	**0.57**	0.21
5011_Cuneus_L	**0.48**	**0.49**	**0.53**	0.14
5012_Cuneus_R	**0.51**	**0.48**	**0.53**	0.11
5021_Lingual_L	**0.42**	**0.52**	**0.43**	0.09
5022_Lingual_R	**0.45**	**0.48**	**0.42**	0.16
5101_Occipital_Sup_L	**0.53**	**0.45**	**0.53**	0.22
5102_Occipital_Sup_R	**0.45**	**0.44**	**0.47**	0.19
5201_Occipital_Mid_L	**0.43**	**0.46**	**0.51**	0.22
5202_Occipital_Mid_R	**0.43**	**0.41**	**0.44**	0.29
5301_Occipital_Inf_L	**0.45**	**0.48**	**0.46**	0.27
5302_Occipital_Inf_R	**0.43**	**0.43**	**0.44**	0.29
5401_Fusiform_L	0.26	0.33	0.32	0.07
5402_Fusiform_R	0.28	0.33	0.33	0.07
6001_Postcentral_L	0.27	0.32	0.30	0.02
6002_Postcentral_R	0.32	0.31	0.31	0.07
6101_Parietal_Sup_L	0.35	0.34	0.37	0.04
6102_Parietal_Sup_R	0.38	0.31	0.40	0.21
6201_Parietal_Inf_L	0.26	0.39	0.33	0.03
6202_Parietal_Inf_R	0.26	**0.41**	**0.43**	0.13
6211_SupraMarginal_L	**0.43**	**0.57**	**0.45**	0.11
6212_SupraMarginal_R	0.31	**0.44**	**0.42**	0.11
6221_Angular_L	0.29	**0.51**	0.37	0.04
6222_Angular_R	0.20	0.38	0.32	0.16
6301_Precuneus_L	0.30	**0.41**	0.34	0.09
6302_Precuneus_R	0.30	0.35	0.33	0.09
6401_Paracentral_Lobule_L	0.33	0.30	0.27	0.04
6402_Paracentral_Lobule_R	0.35	0.31	0.30	0.01

7001_Caudate_L	0.32	0.31	0.31	0.05
7002_Caudate_R	0.26	0.30	0.30	0.03
7011_Putamen_L	0.22	0.31	0.20	-0.01
7012_Putamen_R	0.23	0.29	0.19	-0.07
7021_Pallidum_L	0.26	0.25	0.13	0.07
7022_Pallidum_R	0.20	0.21	0.14	-0.01
7101_Thalamus_L	0.06	0.28	0.13	-0.06
7102_Thalamus_R	0.06	0.34	0.17	-0.03
8101_Heschl_L	**0.52**	**0.60**	0.39	0.22
8102_Heschl_R	**0.52**	**0.61**	**0.41**	0.15
8111_Temporal_Sup_L	**0.59**	**0.64**	**0.54**	0.20
8112_Temporal_Sup_R	**0.56**	**0.62**	**0.54**	0.13
8121_Temporal_Pole_Sup_L	0.21	0.39	0.25	-0.04
8122_Temporal_Pole_Sup_R	0.27	**0.45**	0.31	-0.09
8201_Temporal_Mid_L	**0.43**	**0.51**	**0.47**	0.09
8202_Temporal_Mid_R	0.39	**0.49**	**0.46**	0.07
8211_Temporal_Pole_Mid_L	0.26	**0.41**	0.21	-0.02
8212_Temporal_Pole_Mid_R	0.25	0.31	0.28	-0.03
8301_Temporal_Inf_L	0.13	0.28	0.22	-0.04
8302_Temporal_Inf_R	0.21	0.22	0.28	0.01
9001_Cerebelum_Crus1_L	0.23	0.29	0.19	-0.03
9002_Cerebelum_Crus1_R	0.21	0.21	0.20	0.02
9011_Cerebelum_Crus2_L	0.14	0.26	0.12	-0.15
9012_Cerebelum_Crus2_R	0.04	0.14	0.16	0.00
9021_Cerebelum_3_L	0.08	0.09	-0.04	0.14
9022_Cerebelum_3_R	0.07	0.14	0.06	0.04
9031_Cerebelum_4_5_L	0.14	0.19	0.11	0.03

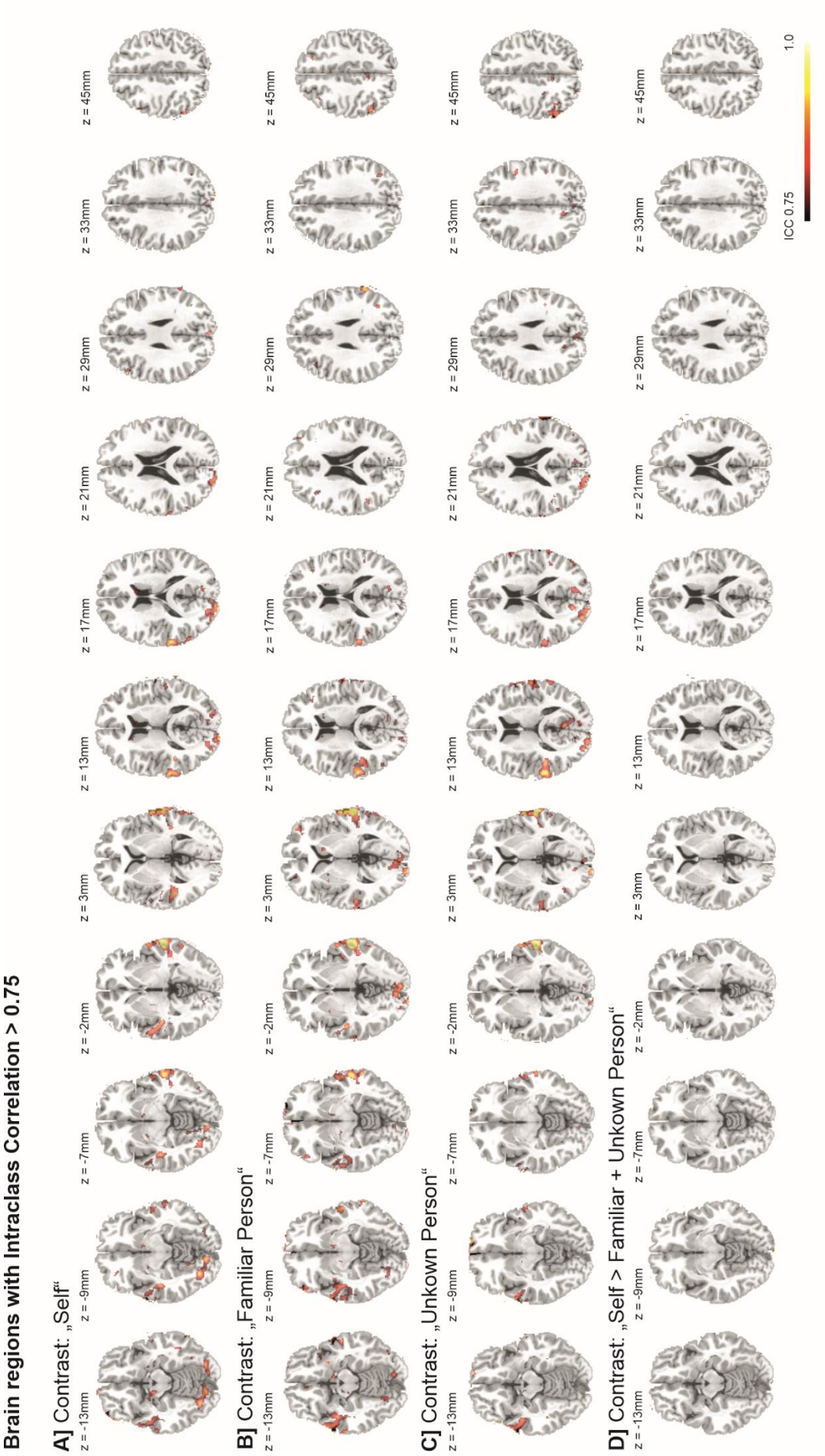
9032_Cerebelum_4_5_R	0.10	0.18	0.16	0.00
9041_Cerebelum_6_L	0.28	0.32	0.22	-0.05
9042_Cerebelum_6_R	0.28	0.27	0.25	-0.05
9051_Cerebelum_7b_L	0.19	0.18	0.08	-0.12
9052_Cerebelum_7b_R	0.20	0.19	0.14	0.16
9061_Cerebelum_8_L	0.22	0.17	0.07	0.10
9062_Cerebelum_8_R	0.30	0.14	0.08	0.14
9071_Cerebelum_9_L	0.27	0.05	0.12	0.11
9072_Cerebelum_9_R	0.26	0.05	0.09	0.10
9081_Cerebelum_10_L	0.18	0.05	-0.03	0.15
9082_Cerebelum_10_R	0.27s	-0.01	0.14	0.15
9100_Vermis_1_2	0.20	0.18	-0.03	0.22
9110_Vermis_3	0.06	0.14	-0.02	0.21
9120_Vermis_4_5	0.09	0.19	0.09	0.00
9130_Vermis_6	0.08	0.22	0.11	-0.11
9140_Vermis_7	0.10	0.24	0.06	-0.12
9150_Vermis_8	0.18	0.22	0.03	0.05
9160_Vermis_9	0.15	0.17	0.04	0.07
9170_Vermis_10	0.32	0.22	0.12	0.22

Bold font = Areas in which mean ICC values exceed the threshold for moderate reliability (ICC > 0.40)

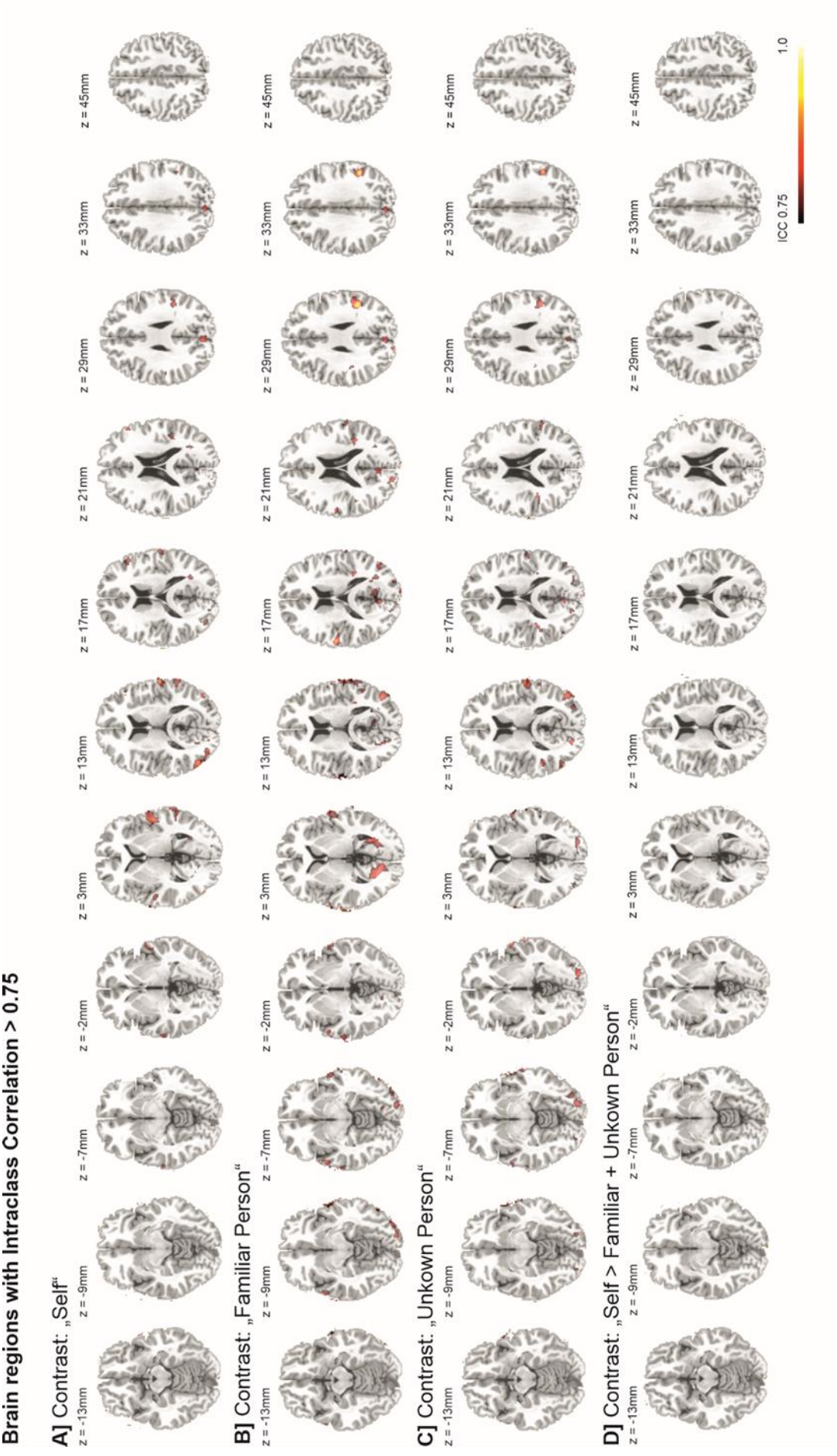
Supplementary Figure S 2.7 Depiction of brain areas that show significant activation in patients and healthy participants for the different task contrasts: "Self", "Familiar Person", "Unknown Person" and "Self > Familiar + Unknown Person" (One-sample t-test, $p_{FWE} < .05$ whole-brain corrected).



Supplementary Figure S 2.8 Depiction of brain areas that show good to excellent reliability (Intra-class correlation [ICC] > 0.75) for the constituent task conditions "self", "familiar person", and "unknown person" as well as the contrast "self > familiar and unknown person" in the patient group (N=11).


Brain regions with Intraclass Correlation > 0.75
A] Contrast: „Self“
B] Contrast: „Familiar Person“
C] Contrast: „Unkown Person“
D] Contrast: „Self > Familiar + Unkown Person“
z = -13mm
z = -9mm
z = -7mm
z = -2mm
z = 3mm
z = 13mm
z = 17mm
z = 21mm
z = 29mm
z = 33mm
z = 45mm
ICC 0.75
1.0

Supplementary Figure S 2.9 Depiction of brain areas that show good to excellent reliability (Intra-class correlation [ICC] > 0.75) for the constituent task conditions "self", "familiar person", and "unknown person" as well as the contrast "self > familiar and unknown person" in the control group (N=29).


Brain regions with Intraclass Correlation > 0.75
A] Contrast: „Self“
B] Contrast: „Familiar Person“
C] Contrast: „Unkown Person“
D] Contrast: „Self > Familiar + Unkown Person“
z = -13mm
z = -9mm
z = -7mm
z = -2mm
z = 3mm
z = 13mm
z = 17mm
z = 21mm
z = 29mm
z = 33mm
z = 45mm
ICC 0.75
1.0

3 DISCUSSION

The results of all three empirical studies consistently show a poor overall global test-retest reliability of the fMRI difference contrast, computed by subtracting two constituent task conditions. This stood in contrast to a moderate to good test-retest reliability of the constituent task conditions underlying the difference score.

3.1 Test-retest reliability of common fMRI block design tasks

With respect to the first research question, the three presented studies showed poor overall reliability of the fMRI difference contrast, computed by subtracting two constituent task conditions, which in the case of the empirical studies were the contrasts "alcohol – neutral", "food – neutral" and "self – other". Consistent with our *first hypothesis*, the reliability estimates for the constituent task conditions (study 1: "alcohol", "neutral"; study 2: "food", "neutral"; study 3: "self", "familiar person", "unknown person") exceeded the reliability of the difference contrast across all empirical studies, showing moderate to good reliability.

3.1.1 Reliability of the difference score

The overall test-retest reliability of the difference contrasts was estimated by computing the Dice and Jaccard coefficients, as well as the ICC and similarity between brain activation maps. The overlap methods indicated that only 1% to 27% of significant voxels that were observed during the first test session remained significantly active during the re-test session, indicating that only a very small fraction of super-threshold voxels showed reliable activation. These low values were observed consistently across the three empirical studies and in all sub-samples. This support the notion that the low Dice and Jaccard coefficients are not due to specifics of the fMRI paradigm, the sample or the duration between test session and re-test session, which were substantially different across the three empirical studies. The mean ICC values across the three empirical studies and all sub-samples fell well below the threshold for moderate test-retest reliability of 0.4, indicating a limited stability of the magnitude of brain activation

across the test session. This was also consistent across all three empirical studies. Considering the results of the similarity analyses, results of the three empirical studies indicated that only about 22% to 73% of the individuals could be re-identified based on the neural activation pattern captured by the difference score.

These results are in line with previous studies. In their meta-analysis, Elliott and colleagues (2020) found mean reliability estimates that resemble the mean reliability, which was found across the three empirical studies, which are presented here. The results of the meta-analysis showed an average ICC of 0.397. This, however, is far below the boundary, which is considered to reflect good reliability (ICC > 0.6). The results also indicated a substantial variability of reliability scores across the different studies. Only a single study reported reliability estimates that fell within the range of good to excellent reliability (ICC > 0.8) (Rath et al., 2016). This study investigated a motor and a sensory fMRI paradigm in seventeen individuals from one day to another.

A subsequent moderator analysis by Elliott et al. (2020) indicated that neither task type (i.e., paradigm); task design (e.g., block vs. event-related); task length; test-retest interval; ROI type (i.e., structural vs. functional); nor sample type (i.e., healthy vs. clinical) significantly moderated the magnitude of the reliability estimates. The authors conducted additional reliability analyses on fMRI data from the Human Connectome Project, arguing that low reliability estimates might have resulted from outdated fMRI scanner hardware or sub-optimal processing pipelines. For this purpose, the authors analyzed reliability in a set of pre-defined ROIs. They found that the test-retest reliability of brain activation in anatomically and functionally defined ROIs was poor (mean $ICC_{anatomical\ ROIs}$ = 0.251 and mean $ICC_{functional\ ROIs}$ = 0.381). The results also indicated that ROIs targeted by the task did not yield higher reliability, compared to "non-target" ROIs. For the reliability analyses of the fMRI data from the Dunedin study, the authors reported similar results with mean reliability below the threshold for moderate reliability.

The results by Elliott et al. (2020) are in line with reports of Schacht and colleagues (Schacht et al., 2011; Schacht et al., 2017). In an early study, they reported good to

excellent test-retest reliability of brain activation (ICC > 0.8) during an alcohol cue-reactivity paradigm over fourteen days in the right VS and DS ROIs in a small sample of nine individuals (Schacht et al., 2011). In contrast, a more recent reliability analysis within the framework of a larger study by this group, using the same fMRI paradigm indicated poor to moderate (ICC < 0.44) reliability of brain activation in the same ROIs over roughly fourteen days (Schacht et al., 2017).

3.1.2 Reliability of the constituent task conditions

Contrary to the poor reliability of the difference contrast across the three empirical studies, the overall test-retest reliability of the constituent task conditions, which underlie the investigated difference contrasts ("alcohol – neutral", "food – neutral" and "self – other"), was moderate to good. Computing the Dice and Jaccard coefficients for the constituent task contrasts (study 1: "alcohol", "neutral"; study 2: "food", "neutral"; study 3: "self", "familiar person", "unknown person"), we found that about 25% to 71% of significant voxels that were observed during the first test session remained significantly active during the re-test session. This indicates that the proportion of significant voxels that remained significantly active between test sessions was up to three times higher for the constituent task conditions, compared to the difference score. The mean ICC values across the three empirical studies showed values in the range between ICC 0.21 and 0.54 and local ICC values > 0.75 in several key regions of the mesocorticolimbic system. The results of the similarity analyses showed values >77% (study 1), > 45% (study 2), and > 36% (study 3) across the three empirical studies, indicating that a higher proportion of participants could be re-identified based on the neural activation patterns captured by the constituent task conditions.

Across the three empirical studies, the reliability estimates indicated moderate to excellent reliability between the first and second fMRI session for all constituent task conditions of the three fMRI paradigms. This supports the general potential of fMRI-based measures to provide reliable indicators of individual brain activation in areas that are part of the addiction network.

This is also in line with the findings by Infantolino and colleagues (2018), who showed excellent reliability of the constituent task conditions of the fMRI faces paradigm ("faces" and "forms"), while the difference contrast ("faces – forms") showed a poor reliability. Like the paradigms that were investigated in the presented empirical studies, this paradigm is a block-design fMRI paradigm with multiple task conditions, which are subtracted from another to build the so-called difference score. This supports the notion that – even though the fMRI-based measures in principle can provide reliable measures – the computation of the difference score leads to an impairment of the reliability of fMRI-based measures.

3.2 Determinants of the reliability of difference scores in fMRI block design tasks

With respect to the second research question, the three presented empirical studies showed a significant and substantial correlation between the constituent task conditions (study 1: "alcohol", "neutral"; study 2: "food", "neutral"; study 3: "self", "familiar person", "unknown person") between r = 0.28 and r = 0.49. Depending on the level of correlation between the constituent task conditions, the resulting reliability estimates of the difference score were lower, when the correlation between the constituent task conditions was high. These findings are in line with *hypothesis 2*.

Several reasons might account for divergence between the moderate to excellent reliability of the constituent task conditions of the three block design tasks and the poor reliability of the difference contrast of the same fMRI tasks. First of all, previous studies demonstrated that – in the case of a substantial correlation between the constituent task conditions of a block design task – a substantial proportion of the shared variance is removed, when subtracting both task conditions (Infantolino et al., 2018). Specifically, the authors compared the internal consistency of the brain activation in the amygdala, captured by the constituent task conditions of an established face-matching block design task ("faces", "shapes") to the internal consistency of the difference contrast ("faces – shapes"). Results demonstrated an excellent internal consistency of the two constituent task conditions, while the internal consistency of the difference contrast was close to zero. The authors also found that the brain activation

in the amygdala, captured by the two constituent task conditions, was highly correlated (r = 0.97) and that the majority of variance (> 90%) was shared between both conditions (Hariri et al., 2002; Infantolino et al., 2018). The low reliability of the difference contrast results, because this shared variance is removed when amygdala activation, captured by the faces contrast, is subtracted from amygdala activation captured by the shapes contrast. In principle, these findings can be applied to other studies using block design fMRI tasks. Based on the findings of Infantolino and colleagues (2018), the reliability of a difference contrast between constituent task conditions should be lower, when brain activation during both task conditions is highly correlated, while the reliability should be higher, when brain activation during both task conditions shows a moderate correlation. This conclusion is in line with previous findings by Luking and colleagues (Luking et al., 2017). In their fMRI study, they investigated the internal consistency of BOLD responses during an fMRI reward task for the task conditions "gain feedback" and "loss feedback", as well as the difference score between both conditions. They found a high internal consistency of both task conditions (Spearman Brown Coefficient > 0.70), while the reliability of the difference score was low (Spearman Brown Coefficient <0.36), which was mirrored by an only modest correlation between the constituting task conditions (Luking et al., 2017). The findings of the presented three empirical studies demonstrate a moderate correlation between the constituting task conditions across all studies (study 1: "alcohol", "neutral"; study 2: "food", "neutral"; study 3: "self", "familiar person", "unknown person") with a shared variance between 8% and 24%. These findings are compatible with the interpretation that reliability of the respective difference contrast is limited by the correlation between the brain activation that is captured by the constituting task conditions and the resulting elimination of substantial parts of the shared variance. Taken together, the results of the three presented empirical studies are in line with *hypothesis 2A*.

It should be noted that the sub-perfect reliability of the constituent task conditions suggest that other factors also influence the observed test-retest reliability. These fac-

tors include limitations inherent to the BOLD signal itself (see section 1.3), factors limiting the SNR (see section 1.3), factors relating to the preprocessing and analysis of the fMRI data and factors relating to the specific sample and task under investigation.

3.2.1 Comparison of the test-retest reliabilities of psychometric scales and the fMRI difference contrasts of cue-reactivity and self-concept paradigms

In the presented empirical studies, we compared the test-retest reliability of the functional brain activation captured by the different tasks in contrast to psychometric measures that conceptually measure resembling constructs (i.e. craving and facets of the self-construct).

With regards to the comparison of the reliability of psychometric scales assessing craving and reliability of brain activation captured by the fMRI food cue-reactivity task, the second presented empirical study provided evidence that the mean global reliability of the fMRI signal, captured by the difference contrast "food – neutral" was lower than the reliability of the psychometric craving scale, while also indicating that the local test-retest reliability of brain activation in several areas of the brain exceeded the test-retest reliability of the psychometric craving scale. This suggests two things. Firstly, and not surprisingly, that the global reliability of brain activation across the whole brain – not limited to brain areas implicated specifically in processing of food cues and food craving (Noori et al., 2012) – is lower than the reliability of psychometric scales that assess the construct of craving. Secondly, the local test-retest reliability of brain activation, captured by the difference contrast "food – neutral", in brain regions that have been shown to be involved in processing of visual food cues and food craving (Noori et al., 2012) outperformed the reliability of the psychometric scale assessing food craving, which was measured concurrently during the fMRI sessions using visual analogue scales. This suggests that local brain activation patterns might indeed provide reliable measures of neural food cue-reactivity.

When we compared the test-retest reliability of psychometric scales assessing different facets of the self-concept to the reliability of brain activation, captured by the fMRI

self-concept paradigm, we found comparable results. Specifically, results of empirical study 3 showed that the mean global reliability of the fMRI signal, captured by the difference contrast "self – unknown and familiar other" was lower than the test-retest reliability of a range of psychometric scales assessing facets of the self-concept, such as self-worth, empathy and social anxiety and social competence deficits. In contrast, the local test-retest reliability of brain activation in several areas of the brain, which were implicated in processing aspects of the self-concept (Hu et al., 2016), exceeded the test-retest reliability of the majority of the psychometric scales. This supports the notion that local brain activation in areas, which are involved in the specific processes that conceptually underlie the task performance, can provide reliable measures of neural self-concept related processes.

These findings are consistent with the results reported by Elliott and colleagues (2020). When analyzing the reliability of brain activation in ROIs that are specifically targeted by the respective fMRI paradigm under investigation (e.g. amygdala in emotion processing fMRI paradigms) they found descriptively higher ICC coefficients for these target ROIs, compared to non-target ROIs for the Human Connectom Project dataset ($ICC_{target\ ROIs} = 0.251$ and $ICC_{non\text{-}target\ ROIs} = 0.239$) and the Dunedin Study dataset ($ICC_{target\ ROIs} = 0.358$ and $ICC_{non\text{-}target\ ROIs} = 0.192$), even though the statistical comparisons did not yield significance.

Taken together, the presented evidence partially supports *hypothesis 2B*, i.e. the reliability of validated psychometric scales is higher than the global test-retest reliability of brain activation across the whole brain, captured by the FMRI difference contrasts, but local reliability in ROIs that are targeted by the specific fMRI task surpass the reliability of psychometric scales.

3.3 Consequences of the reliability of difference scores in fMRI block design tasks

With respect to the third research question, the first presented study compared the correlation between brain activation in the putamen and caudate captured by the "alcohol" and "neutral" task conditions of the fMRI alcohol cue-reactivity task, as well as

between the difference contrast and subjective craving for alcohol, measured using the Obsessive Compulsive Drinking Scale (OCDS). Results show that the significant correlation between brain activation in the caudate, captured by the "alcohol" contrast, and the OCDS could not be replicated, when investigating the association between OCDS and brain activation in the same area, captured by the difference contrast ("alcohol – neutral"). The power to detect even small to medium correlations exceeded 78%. This suggests that this finding was not due to power issues.

The relevance of the reliability of the fMRI signal for establishing correlations with behavioral data has been debated previously. Prominently, Vul and colleagues (2009) pointed to the fact that the reliability of fMRI data limits the strength of correlations, which can be established with any external variable. Our data simulations fortified this idea. The results of the data simulation showed that the correlation between the difference score (modelled according to the actual fMRI data) and the behavioral variable (modelled according to the OCDS data) was lower by about one third, compared to the correlation between the constituting task conditions and the external variable. This is in line with *hypothesis 3*.

Our findings are in line with previous studies investigating the association between ICC and effects sizes (Brown et al., 2011) using a working memory fMRI paradigm. Authors found a curvilinear association between effect size estimates (Cohen's d) and ICC magnitudes, suggesting that the effect sizes, derived from fMRI measures, depend on the reliability of the fMRI data. Taken together, previous research and the presented empirical studies supports the notion that the low reliability of fMRI-based measures can substantially impair the capacity to establish associations between these fMRI measures and external variables, and the capacity to detect (true) effects in the data and estimate effect sizes.

3.4 Critical reflection and limitations of fMRI-based measures

Several conceptual and methodological issues limit the capacity of fMRI to provide reliable and replicable measures.

3.4.1 Non-linearity and susceptibility of the BOLD signal

Even though the BOLD signal has been the basis of fMRI research for decades, it is still debatable whether it can be assumed that the BOLD signal itself is sufficiently stable to provide a reliable measure for brain activation. The BOLD signal itself is a complex physiological response and is viewed as indirect surrogate of activation of neuronal ensembles. It has been shown that the BOLD signal reflects changes in deoxyhemoglobin that are driven by changes in local blood flow and blood oxygenation. These changes in blood flow and oxygenation in turn are coupled to neuronal activity by a process termed neurovascular coupling. Currently, however, there is still a substantial gap in the knowledge regarding the processes, which play a role in the neurovascular coupling, as well as the association between the underlying processes and the BOLD signal. Previous studies have implicated different cellular mechanisms to play a role in functional neurovascular coupling, including contributions of neurons, astrocytes, and pericytes (Hillman, 2014). However, there is still an ongoing debate on the relative contribution of these processes.

Regarding the association between BOLD response and the underlying neural processes, basically all fMRI research assumes a linear relationship between both. However, there is no reason to assume a strict linear neurovascular coupling, as this would also mean that the BOLD signal should not vary as a function of either the strength of the neural response or with previous response history, and that the BOLD signal should grow indefinitely in proportion to the underlying neural activity (Boynton, 2011). This highlights that the assumption of a strict linear association between BOLD response and neural activity is questionable. Nevertheless, early studies, which combined fMRI and electrophysiological recordings, suggested that the BOLD signal reliably follows the spiking activation of local neuronal ensembles (Boynton et al., 1996; Dale & Buckner, 1997). Further studies, however, observed that the response to repeated stimulation was smaller than predicted based on a linear model, and that the BOLD response to very brief stimulation is larger than predicted (Huettel & McCarthy, 2000; Vazquez & Noll, 1998). A more recent study also observed that the synaptic and

spiking responses of neurons were more selective than the vascular responses in the way that small vessels in responded to stimuli that elicited only minimal neural activity in the local tissue. These findings suggest that several aspects of the neurovascular coupling process are non-linear, and that local neural and haemodynamic responses are partly decoupled (O'Herron et al., 2016). Consequently, variability in the BOLD signal does not necessarily reflect changes in neuronal activity, and vice versa. Interestingly, larger deviations from linearity were observed in fMRI studies using more complex fMRI paradigms (Boynton, 2011), while linearity is approximated when simple stimulation fMRI paradigms were used (Boynton et al., 1996; Dale & Buckner, 1997). Against this background, it can be argued that the higher reliability (ICC > 0.80), which was reported for rather simple motor and sensorimotor stimulation fMRI tasks (Rath et al., 2016) could be due to the fact that the neurovascular coupling approximates linearity, thus more closely meeting the assumption of the statistical models used to analyze the data.

Beyond that, several studies have pointed towards a susceptibility of the BOLD signal to endogenous and exogenous influences. Specifically, the age of an individual, blood pressure, hormonal status, body mass index and time of the day were associated with changes in blood flow and hence the BOLD signal (Alosco et al., 2014; Curtis et al., 2016; Muller et al., 2012; Whitworth et al., 2005). All these parameters introduce additional sources of variability to the fMRI BOLD signal and suggest that the BOLD signal might neither be stable within an individual, nor comparable between two individuals that show different properties on the above listed factors. This might also explain why different studies using the same fMRI paradigm yielded different reliability estimates. This strongly supports the argument that a standardization of fMRI studies is necessary, in order to limit the bias that is introduced by the above stated factors. Regarding the alcohol cue-reactivity paradigm, an international consensus group has proposed a checklist, which can be guide the standardization of the fMRI-based measurement of neural cue-reactivity (Ekhtiari et al., 2020). Such efforts are needed to establish standards for collecting, analyzing and interpreting fMRI data, in order to limit the sources of variance.

3.4.2 Parameters of fMRI acquisition and analysis that influence reliability

In principle, all factors limiting the SNR can impair the resulting reliability of fMRI task measures (see section 1.3). In addition, parameters in the analytical pipeline of fMRI data can introduce additional variance in the data that limit the reliability of fMRI-based measures. Different analysis pipelines applied on the same dataset may not produce the same results. Differences in the software versions, operating systems and algorithms can all contribute to variability in results. Recent work analyzed publicly available fMRI datasets using three common software packages, the Analysis of Functional NeuroImages (AFNI) software, the FMRIB Software Library (FSL) and SPM using the parameters of the respective source publication (Bowring et al., 2019). Results showed marked differences, such as Dice similarity coefficients ranging from 0.000 to 0.684, depending on the software package. Across the re-analyzed datasets, the authors reported considerable differences between the AFNI, FSL, and SPM results on all investigated metrics. For example, Bland–Altman plots showed that differences between reported *T*- values across software packages were as large as four for a considerable quantity of voxels. Euler Curves showed a substantial difference in the number of clusters that were found using the separate software packages. The findings show that small effects may not be replicable, when analyzing datasets with different software packages. Authors also concluded that there is currently no gold-standard and reference point when attempting to analyze fMRI datasets and hence no direct comparison can be made between analytical strategies with regards to optimal parameters. The same authors compared the statistical fMRI maps, in order to determine the main sources of variability between the software packages (Bowring et al., 2022). The authors found that the variations between the software packages are largely attributable to a few individual analysis stages, such as the choice of the first-level signal model and first-level noise model. In contrast, group-level results were largely unaffected by which software package is used to model the low-frequency fMRI drifts. These results stress the need for a harmonization of analytical pipelines.

Unfortunately, there is currently no agreed-upon standard in the field. Authors suggested that raw data could be analyzed using a range of workflows. By deriving a range of analysis results from a single dataset, meta-analytic methods could be applied to account for the variability between pipelines and integrate inconsistent findings. In addition, until standardization of analytical pipelines is achieved, replication studies seem important, in order to provide evidence for the robustness of single study fMRI findings.

3.4.3 Limitation of reliability estimates

A range of methods have been developed to estimate the test-retest reliability of fMRI, which capture different aspects of reliability. While several measures address the proportion of voxels that remain significant from one test session to a follow-up test session, other methods assess the similarity of the magnitude of brain activation. To date however, there is no consensus on what method should be used as a standard.

The methods that are most frequently applied by fMRI studies are the ICC and the overlap methods (see also section 1.3.2 for details). However, both methods face several important limitations. For example, the magnitude of the ICC depends on the heterogeneity of the sample under investigation. If all other factors are kept constant, more heterogeneous samples with higher between-subject variance will result in higher ICC values, while more homogenous sample would yield lower ICC values. This also limits the capacity to transfer the findings from one study sample to another study sample. The dependence of the ICC on the between-subject variance can also be applied to the experimental fMRI paradigm under investigation. A fMRI paradigm that produces a higher between-subject variability in neural response patterns would yield higher ICC values (Bennett & Miller, 2010). The overlap methods, such as the Dice and Jaccard coefficients, face major limitations. The first one is the fact that the results obtained by both methods heavily depend on the pre-defined threshold that defines significantly active voxels (e.g. $p < 0.001$). For higher p-values, the overlap methods produce higher values. This can be illustrated when assuming that a p-value of 1.0 is

chosen. In this case, both overlap methods would show a perfect overlap of 100%, because all voxels would be defined as "active", even though the majority would just show noise and falsely active voxels. Hence, this finding would be meaningless. In addition, due to its definition, the Dice coefficient will also yield higher values compared to the Jaccard coefficient, which complicates a direct comparison of both measures. This illustrates that the choice of an optimal p-threshold is not trivial. Nevertheless, there is no established standard on which p-value should be chosen. This complicates the comparability of overlap estimates across studies. Another major limitation to all available reliability estimates for fMRI data is that there is no established optimal threshold to interpret these data. Whilst there have been suggestions on which ICC values could be considered as moderate, good or excellent (Shrout & Fleiss, 1979), the use of these thresholds and their interpretation currently varies substantially across studies. In addition, for other reliability estimates, there is currently no established standard on how to interpret the resulting reliability estimates, which complicates the interpretation and comparability of studies.

Other methods exist to determine the different aspects of fMRI reliability (see also section 1.3.2 for details). However, these methods are less common, and are sometimes complicated to compute and interpret, with no standards having been established as to how to interpret the resulting values regarding which values can be considered to reflect "good" reliability.

In summary, there is a dire need to establish standards regarding which reliability estimates should be used, together with common thresholds to interpret the resulting values in order to advance the efforts of fMRI reliability research.

3.5 Future Directions

3.5.1 Better characterization of the factors that influence reliability

Currently, there is few data on how reliable and replicable the cognitive domains are that are to be assessed using fMRI. Longitudinal studies are required to establish the "natural" course of these cognitive domains and to determine their reproducibility and

reliability. In addition, most neuroimaging studies currently do not report reliability metrics. Even if reliability estimates are reported, the precision of such estimates depends on the sample size. Large sample sizes of 150 or more participants seem necessary in order to provide precise reliability estimates for an fMRI task. In addition, studies should explore moderators of test-retest reliability (e.g., the test-retest interval). Studies that systematically assess the influence of potential moderators are needed to inform future fMRI studies. Currently, there is a lack of knowledge on which factors might moderate test-retest reliability.

Furthermore, the fMRI task design might also influence the test-retest reliability. Currently, most tasks were developed to maximize within-subject differences (i.e. differences between task conditions), instead of between subject differences. Research trying to identify neural biomarkers and associations between neural brain responses and behavior, however, relies on stable between-subject differences and variance. fMRI tasks could be optimized with regards to providing reliable individual-differences measures e.g. by selecting stimuli for fMRI tasks on the basis of their capacity to distinguish between different groups or to elicit reliable between-subjects variance.

The heterogeneity in study design, fMRI acquisition, analytical pipelines and participant characteristics currently limit the reproducibility in the field of fMRI research. A harmonization of task designs, fMRI procedures and analytical steps seems important, in order to enhance comparability between studies and reproducibility. Efforts have been undertaken by several consortia and groups to propose standards or make recommendations on parameters and study details, which should be reported, including sample details, details on the fMRI task and details on the analytical pipeline (Ekhtiari et al., 2020). Overall, harmonization of fMRI methods across studies seems important with regards to enhancing comparability between studies.

3.5.2 Development of novel measures to quantify fMRI signals

A range of different metrics are currently used by different studies to quantify the captured fMRI signal using the respective fMRI task. For example, a range of contrast images can be calculated from the same fMRI task (e.g. difference between to constituent task conditions or one task condition vs. implicit baseline), which most likely produces different results. In addition, the fMRI signal can be expressed using different metrics, such as, but not limited to, the weighted or unweighted mean of activation values in a specific region or the sum of voxels in a brain area that surpass a predefined statistical threshold or the maximum t-statistic value in a specific region or the sum of t-statistics in a region (Reinhard et al., 2015). Previous work has shown that these measures perform very differently with regards to predicting clinical outcome (e.g. relapse risk). At the moment, most metrics are derived from difference contrast maps that reflect the difference of brain activation statistics between two task conditions (e.g. alcohol, neutral). As evidenced by the presented empirical studies and previous work, such difference scores will have lower reliability than the constituent task conditions and thus limit the reliability of task fMRI (Infantolino et al., 2018). Alternative approaches to expressing the magnitude of brain activation which is captured using fMRI might overcome some of the limitations. Multivoxel pattern analysis might be one approach to exploiting the high dimensionality of fMRI data to develop measures with latent-variable models that capture individual differences in representational spaces (Cooper et al., 2019) and confirm reliability through prediction of individual differences in independent samples (Yarkoni & Westfall, 2017). In addition, combining the data of an individual from different fMRI tasks could enhance the precision of the resulting measure by reducing the impact of the measurements error of a single parameter. Related to this issue, currently many fMRI studies use the sample mean as a reference points for interpreting the obtained data, e.g. which participants show a high or low response. As the mean, however, heavily depends on the sample under investigation and may be prone to extreme values, especially in the case of small sam-

ples, a shift to measures that do not rely on a specific sample seems important to enhance comparability, and to allow transfer of results from one study to new datasets. One option with regards to differences between two task conditions could be to define o as the reference point, e.g. when brain activation is similar in both task conditions. This would create a threshold, which is independent from group means and sample homogeneity. However, the meaningfulness of such a threshold relies on the question under investigation. Another option would be the use of individual level comparisons, i.e. increase vs. decrease of brain activation over fMRI task blocks to create an individual-level parameter that is independent from the other study datasets. Again, the meaningfulness of such a parameter needs to be established for the specific question under investigation. Also, with regards to this point, establishing standards across fMRI studies seems important to promote transfer of results from one study to new datasets.

3.5.3 Prospects for translation to clinical care

Even though currently fMRI research is hampered by test-retest reliability, a lack of replication and missing standards for data collection, data processing and data reporting, several lines of evidence support its potential for translation to clinical care. Previous research has shown that predictive models, which incorporate structural and functional MRI data yield a higher accuracy in predicting a subsequent relapse, compared to models that solely rely on clinical patient data (Seo et al., 2015). This finding supports the potential of MRI data to predict clinically relevant patient outcomes. These results were supported by other studies, which repeatedly demonstrated that neural alcohol cue-reactivity predicts relapse risk in alcohol dependent patients after withdrawal treatment (Bach et al., 2015; Reinhard et al., 2015). Further studies also provided evidence that patients that show a high neural response to alcohol cues are those showing a better treatment response to naltrexone, an approved anti-craving medication (Bach, Weil, et al., 2021; K. Mann et al., 2014). Furthermore, studies showed that the neural response to alcohol cues is also sensitive to the effects of anti-relapse medication, such as naltrexone and nalmefene (Bach et al., 2020; Karl et al.,

2021), indicating that neural alcohol cue-reactivity might be capable of monitoring effective treatment. In addition, a reduction in striatal alcohol cue-induced brain activation during pharmacological relapse-prevention treatment with naltrexone been associated with lower relapse risk and better patient outcome (Bach et al., 2020; Schacht et al., 2017). Furthermore, preliminary studies indicated the real-time fMRI can be used to modify the extent of alcohol cue-induced brain activation in key regions of the mesolimbic reward system (Kirsch et al., 2016). Even though the potential of this method for translation into clinical care needs to be confirmed, it suggests that neural alcohol cue-reactivity might not only provide a potential biomarker for relapse risk and successful treatment, but might also represent a target for treatment interventions.

Likewise, studies in obese patients demonstrated a significant relationship between neural food-cue reactivity during the presentation of food stimuli and the extent of subjective craving for food. For example, an early study by Pelchat and colleagues demonstrated significant correlations between neural food-cue reactivity in the hippocampus, insula, caudate, and food craving (Pelchat et al., 2004). Further studies confirmed the significant association between neural food-cue reactivity in the caudate, thalamus, hippocampus, cerebellum, and posterior cingulate with the magnitude of subjective craving for food (Jastreboff et al., 2013).

Three studies involving diet and lifestyle intervention examined the relationship between neural stimulus reactivity at baseline and the degree of success and maintenance of weight loss. These studies showed significant associations between successful weight loss and neural food-cue reactivity in the insula, ACC, NAcc, and ACC (Murdaugh et al., 2012), orbitofrontal cortex (Weygandt et al., 2019), PFC (Nock et al., 2012), and putamen, caudate, and pallidum (Hermann et al., 2019). In a recent study, authors examined the relationship between neural food-cue reactivity before starting a low-calorie diet, and after one, and three months of treatment with treatment success in terms of greater weight loss in obese individuals (Neseliler et al., 2019). Initial weight loss in the first month of treatment correlated positively with an increase in

neural food cue-reactivity in the dlPFC, IFG, dorsal ACC, inferior parietal lobule, and caudate. Food cue-reactivity in these areas also correlated positively with subsequent weight loss from the first to the third month of treatment. In addition, a reduction in neural stimulus reactivity in these areas (i.e., a return to baseline) correlated with a return to weight gain 2 years later. Another study that examined the relationship between neural food cue-reactivity at the start of a weight loss intervention (3-month program consisting of exercise and calorie reduction) and treatment success reported an indirect effect of neural food cue-reactivity on weight loss mediated by treatment adherence (Szabo-Reed et al., 2020).

In addition, studies showed that participants who successfully lost weight and maintained their weight loss had differential neural food cue-reactivity in the parietal cortex (Tregellas et al., 2011) and the PFC and ACC (McCaffery et al., 2009). In a recent longitudinal imaging study from our group that examined obese patients before and after bariatric surgery using fMRI, there was a significant increase in neural food cue-reactivity in the OFC and a decrease in neural stimulus reactivity in the amygdala from the time point before surgery compared to both examinations 8 weeks and 24 weeks after surgery (Bach, Grosshans, Koopmann, Pfeifer, et al., 2021). Here, activation in the amygdala showed a positive correlation with the level of subjective desire for food, whereas activation in the OFC showed a negative correlation with desire for food. Consistent with these findings, another study showed that more successful weight loss after gastric bypass surgery was associated with increased neural food cue-reactivity in the dlPFC (Goldman et al., 2013).

Other studies comparing neural food cue-reactivity before and after bariatric intervention showed similar findings. For example, there was a reduction in neural food cue-reactivity in the insula and putamen after bariatric intervention (Bruce et al., 2012; Ochner et al., 2011; Ochner, Laferrère, et al., 2012; Ochner, Stice, et al., 2012).

In addition to studies on the effect of bariatric intervention on neural food cue-reactivity, some studies also examined the effect of behavioral interventions. For example, a study comparing the effect of caloric restriction via meal replacement with the effect

of caloric restriction by reducing meal size showed that caloric restriction via meal replacement (shakes) resulted in higher neural food cue-reactivity in the dlPFC, OFC, ACC, insula, and NAcc (Kahathuduwa et al., 2018).

A preliminary study examined the effect of neurofeedback on neural food cue-reactivity during the presentation of food stimuli in normal weight subjects (Ihssen et al., 2017). Neurofeedback training aimed to help subjects learn to downregulate their brain activation during exposure to appetitive images of food stimuli. The fMRI feedback was provided by reducing or increasing the size of the food images. The authors reported a reduction in neural food cue-reactivity in the amygdala, insula, dmPFC, and precuneus and cuneus following neurofeedback training. In addition, the authors reported a significant reduction in subjective hunger sensation after neurofeedback intervention, which correlated with neural food cue-reactivity in the amygdala.

Taken together, the three empirical studies and previous research point towards the potential of neural brain activation, captured using fMRI, as markers (e.g. for predicting treatment responses) and also as targets for non-pharmacological and pharmacological interventions. However, to exploit the full potential of fMRI-based measures, a standardization with regards to data collection, data analysis and data interpretation seems to be important and needs to be established in the future to confirm the robustness of single-study findings.

4 SUMMARY

The identification of "neural biomarkers" of psychiatric disorders has been the aim of many recent fMRI studies. This endeavor, however, critically depends on the test-retest reliability of the corresponding fMRI task under investigation, which is an essential prerequisite for establishing associations between neural brain activation patterns and behavior. Previous research highlighted the low overall reliability of fMRI-based measures across studies and pointed towards the negative impact that might arise from computing difference scores between two correlated constituent task conditions, when applying fMRI tasks (Elliott et al., 2020; Infantolino et al., 2018). The presented thesis investigated whether the test-retest reliability of the constituent conditions of an fMRI block design paradigm is higher than the reliability of the difference score, which is created by subtracting brain activation during one condition from another condition and is commonly used across a range of block design fMRI paradigms. All three empirical studies show that the reliability estimates of the constituent task conditions across three different block-design fMRI paradigms exceed the reliability estimates of the difference score. In addition, the presented empirical studies show that the level of correlation between the constituent conditions of the respective fMRI paradigm determined the level of the resulting reliability of the difference score.

While moderate to good reliability of the difference score could be demonstrated in several brain regions and comparison of the test-retest reliability of the fMRI signal to the test-retest reliability of psychometric scales used to assess e.g. craving for food, in principle support the potential of fMRI of providing reliable measures, the presented results also challenge the application of difference scores when analyzing fMRI block-design tasks. This seems especially important in the light of the results of the first empirical study that demonstrated that the inter-correlation between the constituent task conditions of a difference score determine the magnitude of associations between the fMRI signal and external variables, and the precision with which effect size estimates can be derived from a specific fMRI study.

A part of the limitations of fMRI-based measures that were identified by the presented empirical studies could be overcome in the future by establishing standards for fMRI data collection, data analysis and data interpretation, as well as using individual-level metrics to express brain activation, and by exploration and consideration of factors that influence test-retest reliability of fMRI block-design studies.

5 REFERENCES

Alosco, M. L., Brickman, A. M., Spitznagel, M. B., Narkhede, A., Griffith, E. Y., Raz, N., . . . Gunstad, J. (2014). Higher BMI is associated with reduced brain volume in heart failure. *BMC Obes*, *1*(1), 4. https://doi.org/10.1186/2052-9538-1-4

Anton, R. F., Moak, D. H., & Latham, P. (1995). The Obsessive Compulsive Drinking Scale: a self-rated instrument for the quantification of thoughts about alcohol and drinking behavior [Comparative Study

Research Support, U.S. Gov't, P.H.S.]. *Alcohol Clin Exp Res*, *19*(1), 92-99. http://www.ncbi.nlm.nih.gov/pubmed/7771669

Association, A. P. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5™, 5th ed* [doi:10.1176/appi.books.9780890425596]. American Psychiatric Publishing, Inc. https://doi.org/10.1176/appi.books.9780890425596

Bach, P., Grosshans, M., Koopmann, A., Kienle, P., Vassilev, G., Otto, M., . . . Kiefer, F. (2021). Reliability of neural food cue-reactivity in participants with obesity undergoing bariatric surgery: a 26-week longitudinal fMRI study. *Eur Arch Psychiatry Clin Neurosci*, *271*(5), 951-962. https://doi.org/10.1007/s00406-020-01218-8

Bach, P., Grosshans, M., Koopmann, A., Pfeifer, A. M., Vollstadt-Klein, S., Otto, M., . . . Kiefer, F. (2021). Predictors of weight loss in participants with obesity following bariatric surgery - A prospective longitudinal fMRI study. *Appetite*, *163*, 105237. https://doi.org/10.1016/j.appet.2021.105237

Bach, P., Hill, H., Reinhard, I., Gädeke, T., Kiefer, F., & Leménager, T. (2021). Reliability of the fMRI-based assessment of self-evaluation in individuals with internet gaming disorder. *Eur Arch Psychiatry Clin Neurosci*. https://doi.org/10.1007/s00406-021-01307-2

Bach, P., Reinhard, I., Koopmann, A., Bumb, J. M., Sommer, W. H., Vollstädt-Klein, S., & Kiefer, F. (2022). Test-retest reliability of neural alcohol cue-reactivity: Is there light at the end of the magnetic resonance imaging tube? *Addict Biol*, *27*(1), e13069. https://doi.org/10.1111/adb.13069

Bach, P., Vollstädt-Klein, S., Kirsch, M., Hoffmann, S., Jorde, A., Frank, J., . . . Kiefer, F. (2015). Increased mesolimbic cue-reactivity in carriers of the mu-opioid-receptor gene OPRM1 A118G polymorphism predicts drinking outcome: a functional imaging study in alcohol dependent subjects. *Eur Neuropsychopharmacol*, *25*(8), 1128-1135. https://doi.org/10.1016/j.euroneuro.2015.04.013

Bach, P., Weil, G., Pompili, E., Hoffmann, S., Hermann, D., Vollstadt-Klein, S., . . . Sommer, W. H. (2021). FMRI-based prediction of naltrexone response in alcohol use disorder: a replication study. *Eur Arch Psychiatry Clin Neurosci*, *271*(5), 915-927. https://doi.org/10.1007/s00406-021-01259-7

Bach, P., Weil, G., Pompili, E., Hoffmann, S., Hermann, D., Vollstadt-Klein, S., . . . Sommer, W. H. (2020). Incubation of neural alcohol cue reactivity after

withdrawal and its blockade by naltrexone. *Addict Biol*, *25*(1), e12717. https://doi.org/10.1111/adb.12717

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of autism and developmental disorders*, *34*(2), 163-175.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Arch Gen Psychiatry*, *4*, 561-571. http://www.ncbi.nlm.nih.gov/pubmed/13688369

Benarroch, E. E. (2010). Neural control of feeding behavior: Overview and clinical correlations. *Neurology*, *74*(20), 1643-1650. https://doi.org/10.1212/WNL.0b013e3181df0a3f

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

Bennett, C. M., Wolford, G. L., & Miller, M. B. (2009). The principled control of false positives in neuroimaging [Research Support, U.S. Gov't, Non-P.H.S.]. *Soc Cogn Affect Neurosci*, *4*(4), 417-422. https://doi.org/10.1093/scan/nsp053

Bhaskaran, K., Douglas, I., Forbes, H., dos-Santos-Silva, I., Leon, D. A., & Smeeth, L. (2014). Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5.24 million UK adults. *Lancet*, *384*(9945), 755-765. https://doi.org/10.1016/S0140-6736(14)60892-8

Bohn, M. J., Krahn, D. D., & Staehler, B. A. (1995). Development and initial validation of a measure of drinking urges in abstinent alcoholics. *Alcohol Clin Exp Res*, *19*(3), 600-606.

Boswell, R. G., & Kober, H. (2016). Food cue reactivity and craving predict eating and weight gain: a meta-analytic review. *Obes Rev*, *17*(2), 159-177. https://doi.org/10.1111/obr.12354

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., . . . Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84-88. https://doi.org/10.1038/s41586-020-2314-9

Bowring, A., Maumet, C., & Nichols, T. E. (2019). Exploring the impact of analysis software on task fMRI results. *Hum Brain Mapp*, *40*(11), 3362-3384. https://doi.org/https://doi.org/10.1002/hbm.24603

Bowring, A., Nichols, T. E., & Maumet, C. (2022). Isolating the sources of pipeline-variability in group-level task-fMRI results. *Hum Brain Mapp*, *43*(3), 1112-1128. https://doi.org/https://doi.org/10.1002/hbm.25713

Boynton, G. M. (2011). Spikes, BOLD, attention, and awareness: a comparison of electrophysiological and fMRI signals in V1. *J Vis*, *11*(5), 12. https://doi.org/10.1167/11.5.12

Boynton, G. M., Engel, S. A., Glover, G. H., & Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. *J Neurosci*, *16*(13), 4207-4221. https://doi.org/10.1523/jneurosci.16-13-04207.1996

Brown, G. G., Mathalon, D. H., Stern, H., Ford, J., Mueller, B., Greve, D. N., . . . Potkin, S. G. (2011). Multisite reliability of cognitive BOLD data. *Neuroimage*, *54*(3), 2163-2175. https://doi.org/10.1016/j.neuroimage.2010.09.076

Bruce, J. M., Hancock, L., Bruce, A., Lepping, R. J., Martin, L., Lundgren, J. D., . . . Savage, C. R. (2012). Changes in brain activation to food pictures after adjustable gastric banding. *Surg Obes Relat Dis*, *8*(5), 602-608. https://doi.org/10.1016/j.soard.2011.07.006

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365-376.

Chiou, J.-s., & Spreng, R. A. (1996). The reliability of difference scores: A re-examination. *Journal of Consumer Satisfaction Dissatisfaction and Complaining Behavior*, *9*, 158-167.

Choi, E. J., Taylor, M. J., Hong, S.-B., Kim, C., Kim, J.-W., McIntyre, R. S., & Yi, S.-H. (2018). Gaming-addicted teens identify more with their cyber-self than their own self: Neural evidence. *Psychiatry Research: Neuroimaging*, *279*, 51-59.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Cooper, S. R., Jackson, J. J., Barch, D. M., & Braver, T. S. (2019). Neuroimaging of individual differences: A latent variable modeling perspective. *Neuroscience & Biobehavioral Reviews*, *98*, 29-46. https://doi.org/https://doi.org/10.1016/j.neubiorev.2018.12.022

Corte, C., & Zucker, R. A. (2008). Self-concept disturbances: Cognitive vulnerability for early drinking and early drunkenness in adolescents at high risk for alcohol problems. *Addict Behav*, *33*(10), 1282-1290.

Curtis, B. J., Williams, P. G., Jones, C. R., & Anderson, J. S. (2016). Sleep duration and resting fMRI functional connectivity: examination of short sleepers with and without perceived daytime dysfunction. *Brain Behav*, *6*(12), e00576. https://doi.org/10.1002/brb3.576

Dale, A. M., & Buckner, R. L. (1997). Selective averaging of rapidly presented individual trials using fMRI. *Hum Brain Mapp*, *5*(5), 329-340. https://doi.org/10.1002/(sici)1097-0193(1997)5:5<329::Aid-hbm1>3.0.Co;2-5

Downey, L., Rosengren, D. B., & Donovan, D. M. (2000). To thine own self be true: Self-concept and motivation for abstinence among substance abusers. *Addict Behav*, *25*(5), 743-757.

Drew Sayer, R., Tamer, G. G., Jr., Chen, N., Tregellas, J. R., Cornier, M. A., Kareken, D. A., . . . Campbell, W. W. (2016). Reproducibility assessment of brain responses to visual food stimuli in adults with overweight and obesity. *Obesity (Silver Spring)*, *24*(10), 2057-2063. https://doi.org/10.1002/oby.21603

Drew Sayer, R., Tamer Jr, G. G., Chen, N., Tregellas, J. R., Cornier, M. A., Kareken, D. A., . . . Campbell, W. W. (2016). Reproducibility assessment of brain responses to visual food stimuli in adults with overweight and obesity. *Obesity*, *24*(10), 2057-2063.

Duncan, K. J., Pattamadilok, C., Knierim, I., & Devlin, J. T. (2009). Consistency and variability in functional localisers. *Neuroimage*, *46*(4), 1018-1026.

Ekhtiari, H., Zare-Bidoky, M., Sangchooli, A., Janes, A. C., Kaufman, M. J., Oliver, J., . . . Zilverstand, A. (2020). A Methodological Checklist for fMRI Drug Cue Reactivity Studies: Development and Expert Consensus. *medRxiv*, 2020.2010.2017.20214304. https://doi.org/10.1101/2020.10.17.20214304

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc Natl Acad Sci U S A*, *113*(28), 7900-7905. https://doi.org/10.1073/pnas.1602413113

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., . . . Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science*, *31*(7), 792-806. https://doi.org/10.1177/0956797620916786

Ernst, T., & Hennig, J. (1994). Observation of a fast response in functional MR. *Magn Reson Med*, *32*(1), 146-149. http://www.ncbi.nlm.nih.gov/pubmed/8084231

Fagerstrom, K. (2012). Determinants of tobacco use and renaming the FTND to the Fagerstrom Test for Cigarette Dependence. *Nicotine Tob Res*, *14*(1), 75-78. https://doi.org/10.1093/ntr/ntr137

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., . . . Constable, R. T. (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat Neurosci*, *18*(11), 1664-1671. https://doi.org/10.1038/nn.4135

Fleiss, J. L. (1986). The Design and Analysis of Clinical Experiments. John Wiley & Sons. *New York*, 11-12.

Fossati, P., Hevenor, S. J., Lepage, M., Graham, S. J., Grady, C., Keightley, M. L., . . . Mayberg, H. (2004). Distributed self in episodic memory: neural correlates of successful retrieval of self-encoded positive and negative personality traits. *Neuroimage*, *22*(4), 1596-1604. https://doi.org/10.1016/j.neuroimage.2004.03.034

Fried, P. J., Jannati, A., Davila-Pérez, P., & Pascual-Leone, A. (2017). Reproducibility of single-pulse, paired-pulse, and intermittent theta-burst TMS measures in healthy aging, type-2 diabetes, and Alzheimer's disease. *Frontiers in aging neuroscience*, *9*, 263.

Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. *Psychological bulletin*, *70*(4), 245.

Frohner, J. H., Teckentrup, V., Smolka, M. N., & Kroemer, N. B. (2019). Addressing the reliability fallacy in fMRI: Similar group effects may arise from unreliable individual effects. *Neuroimage*, *195*, 174-189. https://doi.org/10.1016/j.neuroimage.2019.03.053

Frostig, R. D., Lieke, E. E., Ts'o, D. Y., & Grinvald, A. (1990). Cortical functional architecture and local coupling between neuronal activity and the microcirculation revealed by in vivo high-resolution optical imaging of intrinsic signals [Research Support, Non-U.S. Gov't

Research Support, U.S. Gov't, P.H.S.]. *Proc Natl Acad Sci U S A*, *87*(16), 6082-6086. http://www.ncbi.nlm.nih.gov/pubmed/2117272

Gearhardt, A. N., Corbin, W. R., & Brownell, K. D. (2009). Preliminary validation of the Yale Food Addiction Scale. *Appetite*, *52*(2), 430-436. https://doi.org/10.1016/j.appet.2008.12.003

Goldin, P. R., Ziv, M., Jazaieri, H., Weeks, J., Heimberg, R. G., & Gross, J. J. (2014). Impact of cognitive-behavioral therapy for social anxiety disorder on the neural bases of emotional reactivity to and regulation of social evaluation. *Behav Res Ther*, *62*, 97-106. https://doi.org/10.1016/j.brat.2014.08.005

Goldman, R. L., Canterberry, M., Borckardt, J. J., Madan, A., Byrne, T. K., George, M. S., . . . Hanlon, C. A. (2013). Executive control circuitry differentiates degree of success in weight loss following gastric-bypass surgery. *Obesity (Silver Spring)*, *21*(11), 2189-2196. https://doi.org/10.1002/oby.20575

Goldstein, R. Z., & Volkow, N. D. (2011). Dysfunction of the prefrontal cortex in addiction: neuroimaging findings and clinical implications. *Nat Rev Neurosci*, *12*(11), 652-669. https://doi.org/10.1038/nrn3119

Grosshans, M., Vollmert, C., Vollstadt-Klein, S., Tost, H., Leber, S., Bach, P., . . . Kiefer, F. (2012). Association of leptin with food cue-induced activation in human reward pathways. *Arch Gen Psychiatry*, *69*(5), 529-537. https://doi.org/10.1001/archgenpsychiatry.2011.1586

Grüsser, S. M., Wrase, J., Klein, S., Hermann, D., Smolka, M. N., Ruf, M., . . . Braus, D. F. (2004). Cue-induced activation of the striatum and medial prefrontal cortex is associated with subsequent relapse in abstinent alcoholics. *Psychopharmacology (Berl)*, *175*(3), 296-302.

Harding, I. H., Andrews, Z. B., Mata, F., Orlandea, S., Martínez-Zalacaín, I., Soriano-Mas, C., . . . Verdejo-Garcia, A. (2018). Brain substrates of unhealthy versus healthy food choices: influence of homeostatic status and body mass index. *International Journal of Obesity*, *42*(3), 448-454. https://doi.org/10.1038/ijo.2017.237

Hariri, A. R., Tessitore, A., Mattay, V. S., Fera, F., & Weinberger, D. R. (2002). The amygdala response to emotional stimuli: a comparison of faces and scenes. *Neuroimage*, *17*(1), 317-323. https://www.ncbi.nlm.nih.gov/pubmed/12482086

Hautzinger, M. K., F.; Kühner, C.; Beck, A.T. (2009). *Beck Depressions-Inventar : BDI II*. Frankfurt am Main : Pearson Assessment.

Heatherton, T. F., Kozlowski, L. T., Frecker, R. C., & Fagerstrom, K. O. (1991). The Fagerstrom Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire. *Br J Addict*, *86*(9), 1119-1127. http://www.ncbi.nlm.nih.gov/pubmed/1932883

Heilig, M., Augier, E., Pfarr, S., & Sommer, W. H. (2019). Developing neuroscience-based treatments for alcohol addiction: A matter of choice? *Transl Psychiatry*, *9*(1), 255. https://doi.org/10.1038/s41398-019-0591-6

Heilig, M., Sommer, W. H., & Spanagel, R. (2016). The Need for Treatment Responsive Translational Biomarkers in Alcoholism Research. *Curr Top Behav Neurosci*, *28*, 151-171. https://doi.org/10.1007/7854_2015_5006

Helzer, J. E., Kraemer, H. C., & Krueger, R. F. (2006). The feasibility and need for dimensional psychiatric diagnoses. *Psychol Med*, *36*(12), 1671-1680. https://doi.org/10.1017/s003329170600821x

Heni, M., Kullmann, S., Veit, R., Ketterer, C., Frank, S., Machicao, F., . . . Fritsche, A. (2014). Variation in the obesity risk gene FTO determines the postprandial cerebral processing of food stimuli in the prefrontal cortex. *Molecular Metabolism*, *3*(2), 109-113.

Hermann, P., Gál, V., Kóbor, I., Kirwan, C. B., Kovács, P., Kitka, T., . . . Vidnyánszky, Z. (2019). Efficacy of weight loss intervention can be predicted based on early alterations of fMRI food cue reactivity in the striatum. *Neuroimage Clin*, *23*, 101803. https://doi.org/10.1016/j.nicl.2019.101803

Hill-Bowen, L. D., Riedel, M. C., Poudel, R., Salo, T., Flannery, J. S., Camilleri, J. A., . . . Sutherland, M. T. (2020). The cue-reactivity paradigm: An ensemble of networks driving attention and cognition when viewing drug-related and natural-reward stimuli. *bioRxiv*, 2020.2002.2026.966549. https://doi.org/10.1101/2020.02.26.966549

Hillman, E. M. (2014). Coupling mechanism and significance of the BOLD signal: a status report. *Annu Rev Neurosci*, *37*, 161-181. https://doi.org/10.1146/annurev-neuro-071013-014111

Hoenig, K., Kuhl, C. K., & Scheef, L. (2005). Functional 3.0-T MR assessment of higher cognitive function: are there advantages over 1.5-T imaging? *Radiology*, *234*(3), 860-868. https://doi.org/10.1148/radiol.2343031565

Holsen, L. M., Davidson, P., Cerit, H., Hye, T., Moondra, P., Haimovici, F., . . . Stoeckel, L. E. (2018). Neural predictors of 12-month weight loss outcomes following bariatric surgery. *Int J Obes (Lond)*, *42*(4), 785-793. https://doi.org/10.1038/ijo.2017.190

Hu, C., Di, X., Eickhoff, S. B., Zhang, M., Peng, K., Guo, H., & Sui, J. (2016). Distinct and common aspects of physical and psychological self-representation in the brain: A meta-analysis of self-bias in facial and self-referential judgements. *Neuroscience & Biobehavioral Reviews*, *61*, 197-207.

Huerta, C. I., Sarkar, P. R., Duong, T. Q., Laird, A. R., & Fox, P. T. (2014). Neural bases of food perception: coordinate-based meta-analyses of neuroimaging studies in multiple modalities. *Obesity (Silver Spring)*, *22*(6), 1439-1446. https://doi.org/10.1002/oby.20659

Huettel, S. A., & McCarthy, G. (2000). Evidence for a refractory period in the hemodynamic response to visual stimuli as measured by MRI. *Neuroimage*, *11*(5 Pt 1), 547-553. https://doi.org/10.1006/nimg.2000.0553

Hunter, J. E., & Schmidt, F. L. (1994). Correcting for sources of artificial variation across studies. *The Handbook of Research Synthesis*, *236*, 323.

Ihssen, N., Sokunbi, M. O., Lawrence, A. D., Lawrence, N. S., & Linden, D. E. J. (2017). Neurofeedback of visual food cue reactivity: a potential avenue to alter incentive sensitization and craving. *Brain Imaging Behav*, *11*(3), 915-924. https://doi.org/10.1007/s11682-016-9558-x

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

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., . . . Wang, P. (2010). Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. In: Am Psychiatric Assoc.

Israelashvili, M., Kim, T., & Bukobza, G. (2012). Adolescents' over-use of the cyber world–Internet addiction or identity exploration? *Journal of adolescence*, *35*(2), 417-424.

Jabbi, M., Keysers, C., Singer, T., & Stephan, K. E. (2011). Response to "Voodoo Correlations in Social Neuroscience" by Vul et al.–summary information for the press. *Online: http://www. bcn-nic. nl/replyVul. pdf (Stand 05.01. 2011)*.

Jaccard, P. (1902). Lois de distribution florale dans la zone alpine. *Bulletin de la Société vaudoise des sciences naturelles*, *38*, 69-130. https://doi.org/10.5169/seals-266762

Jastreboff, A. M., Sinha, R., Lacadie, C., Small, D. M., Sherwin, R. S., & Potenza, M. N. (2013). Neural Correlates of Stress- and Food Cue–Induced Food Craving in Obesity. *Association with insulin levels*, *36*(2), 394-402. https://doi.org/10.2337/dc12-1112

Johnson, S. C., Baxter, L. C., Wilder, L. S., Pipe, J. G., Heiserman, J. E., & Prigatano, G. P. (2002). Neural correlates of self-reflection. *Brain*, *125*(Pt 8), 1808-1814.

Kahathuduwa, C. N., Davis, T., O'Boyle, M., Boyd, L. A., Chin, S. H., Paniukov, D., & Binks, M. (2018). Effects of 3-week total meal replacement vs. typical food-based diet on human brain functional magnetic resonance imaging food-cue reactivity and functional connectivity in people with obesity. *Appetite*, *120*, 431-441. https://doi.org/10.1016/j.appet.2017.09.025

Karl, D., Bumb, J. M., Bach, P., Dinter, C., Koopmann, A., Hermann, D., . . . Vollstädt-Klein, S. (2021). Nalmefene attenuates neural alcohol cue-reactivity in the ventral striatum and subjective alcohol craving in patients with alcohol use disorder. *Psychopharmacology (Berl)*. https://doi.org/10.1007/s00213-021-05842-7

Karra, E., O'Daly, O. G., Choudhury, A. I., Yousseif, A., Millership, S., Neary, M. T., . . . Hess, M. E. (2013). A link between FTO, ghrelin, and impaired brain food-cue responsivity. *The Journal of Clinical Investigation*, *123*(8), 3539-3551.

Kerem, L., Hadjikhani, N., Holsen, L., Lawson, E. A., & Plessow, F. (2019). Oxytocin reduces the functional connectivity between brain regions involved in eating behavior in men with overweight and obesity. *International Journal of Obesity*. https://doi.org/10.1038/s41366-019-0489-7

Kiefer, F., Kirsch, M., Bach, P., Hoffmann, S., Reinhard, I., Jorde, A., . . . Vollstadt-Klein, S. (2015). Effects of D-cycloserine on extinction of mesolimbic cue reactivity in alcoholism: a randomized placebo-controlled trial. *Psychopharmacology (Berl)*, *232*(13), 2353-2362. https://doi.org/10.1007/s00213-015-3882-5

Kim, M.-K., Jung, Y. H., Kyeong, S., Shin, Y.-B., Kim, E., & Kim, J.-J. (2018). Neural correlates of distorted self-concept in individuals with internet gaming disorder: a functional MRI study. *Frontiers in psychiatry*, *9*, 330.

Kirsch, M., Gruber, I., Ruf, M., Kiefer, F., & Kirsch, P. (2016). Real-time functional magnetic resonance imaging neurofeedback can reduce striatal cue-reactivity to alcohol stimuli. *Addict Biol*, *21*(4), 982-992. https://doi.org/10.1111/adb.12278

Kivlahan, D. R., Sher, K. J., & Donovan, D. M. (1989). The Alcohol Dependence Scale: a validation study among inpatient alcoholics. *J Stud Alcohol*, *50*(2), 170-175. http://www.ncbi.nlm.nih.gov/pubmed/2927131

Kolbeck, S., & Maß, R. (2009). *Fragebogen zu sozialer Angst und sozialen Kompetenzdefiziten: SASKO*. Hogrefe.

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., . . . Zimmerman, M. (2017). The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *J Abnorm Psychol*, *126*(4), 454-477. https://doi.org/10.1037/abn0000258

Kroemer, N. B., Guevara, A., Vollstädt-Klein, S., & Smolka, M. N. (2013). Nicotine alters food-cue reactivity via networks extending from the hypothalamus. *Neuropsychopharmacology*, *38*(11), 2307-2314. https://doi.org/10.1038/npp.2013.133

Lang, P., Bradley, M., & Cuthbert, B. (1999). The International Affective Picture System (IAPS). In C. f. t. S. o. E. a. Attention (Ed.). Gainesville: University of Florida.

Lee, A. T., Glover, G. H., & Meyer, C. H. (1995). Discrimination of large venous vessels in time-course spiral blood-oxygen-level-dependent magnetic-resonance functional neuroimaging [Research Support, Non-U.S. Gov't

Research Support, U.S. Gov't, P.H.S.]. *Magn Reson Med*, *33*(6), 745-754. http://www.ncbi.nlm.nih.gov/pubmed/7651109

Leménager, T., Dieter, J., Hill, H., Koopmann, A., Reinhard, I., Sell, M., . . . Mann, K. (2014). Neurobiological correlates of physical self-concept and self-identification with avatars in addicted players of Massively Multiplayer Online Role-Playing Games (MMORPGs). *Addict Behav*, *39*(12), 1789-1797.

Leménager, T., Hoffmann, S., Dieter, J., Reinhard, I., Mann, K., & Kiefer, F. (2018). The links between healthy, problematic, and addicted Internet use regarding comorbidities and self-concept-related characteristics. *Journal of Behavioral Addictions*, *7*(1), 31-43.

Lemenager, T., Neissner, M., Sabo, T., Mann, K., & Kiefer, F. (2020). "Who am i" and "how should i be": a systematic review on self-concept and avatar identification in gaming disorder. *Current Addiction Reports*, *7*(2), 166-193.

Lemmens, J. S., Valkenburg, P. M., & Gentile, D. A. (2015). The Internet gaming disorder scale. *Psychological assessment*, *27*(2), 567.

Li, G., Ji, G., Hu, Y., Liu, L., Jin, Q., Zhang, W., . . . Wang, G.-J. (2019). Reduced plasma ghrelin concentrations are associated with decreased brain reactivity to food

cues after laparoscopic sleeve gastrectomy. *Psychoneuroendocrinology*, *100*, 229-236. https://doi.org/10.1016/j.psyneuen.2018.10.022

Loeber, S., Kiefer, F., Wagner, F., Mann, K., & Croissant, B. (2009). [Treatment outcome after inpatient alcohol withdrawal: impact of motivational interventions: a comparative study]. *Nervenarzt*, *80*(9), 1085-1092. https://doi.org/10.1007/s00115-009-2724-2 (Behandlungserfolg nach qualifiziertem Alkoholentzug: Vergleichsstudie zum Einfluss motivationaler Interventionen.)

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal [Research Support, Non-U.S. Gov't]. *Nature*, *412*(6843), 150-157. https://doi.org/10.1038/35084005

Lou, H. C., Luber, B., Crupain, M., Keenan, J. P., Nowak, M., Kjaer, T. W., . . . Lisanby, S. H. (2004). Parietal cortex and representation of the mental self. *Proceedings of the National Academy of Sciences*, *101*(17), 6827-6832.

Luking, K. R., Nelson, B. D., Infantolino, Z. P., Sauder, C. L., & Hajcak, G. (2017). Internal Consistency of Functional Magnetic Resonance Imaging and Electroencephalography Measures of Reward in Late Childhood and Early Adolescence. *Biological psychiatry. Cognitive neuroscience and neuroimaging*, *2*(3), 289-297. https://doi.org/10.1016/j.bpsc.2016.12.004

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*(1), 124-135.

Mann, K., Vollstadt-Klein, S., Reinhard, I., Lemenager, T., Fauth-Buhler, M., Hermann, D., . . . Smolka, M. N. (2014). Predicting naltrexone response in alcohol-dependent patients: the contribution of functional magnetic resonance imaging [Clinical Trial

Research Support, Non-U.S. Gov't]. *Alcohol Clin Exp Res*, *38*(11), 2754-2762. https://doi.org/10.1111/acer.12546

Mann, K., Vollstädt-Klein, S., Reinhard, I., Leménager, T., Fauth-Bühler, M., Hermann, D., . . . Smolka, M. N. (2014). Predicting Naltrexone Response in Alcohol-Dependent Patients: The Contribution of Functional Magnetic Resonance Imaging. *Alcoholism: Clinical and Experimental Research*, *38*(11), 2754-2762. https://doi.org/10.1111/acer.12546

Manuck, S. B., Brown, S. M., Forbes, E. E., & Hariri, A. R. (2007). Temporal stability of individual differences in amygdala reactivity. *Am J Psychiatry*, *164*(10), 1613-1614. https://doi.org/10.1176/appi.ajp.2007.07040609

McCaffery, J. M., Haley, A. P., Sweet, L. H., Phelan, S., Raynor, H. A., Del Parigi, A., . . . Wing, R. R. (2009). Differential functional magnetic resonance imaging response to food pictures in successful weight-loss maintainers relative to normal-weight and obese controls. *Am J Clin Nutr*, *90*(4), 928-934. https://doi.org/10.3945/ajcn.2009.27924

Moser, E., Teichtmeister, C., & Diemling, M. (1996). Reproducibility and postprocessing of gradient-echo functional MRI to improve localization of brain

activity in the human visual cortex. *Magn Reson Imaging*, *14*(6), 567-579. https://doi.org/10.1016/0730-725x(96)00095-1

Muller, M., van der Graaf, Y., Visseren, F. L., Mali, W. P., & Geerlings, M. I. (2012). Hypertension and longitudinal changes in cerebral blood flow: the SMART-MR study. *Ann Neurol*, *71*(6), 825-833. https://doi.org/10.1002/ana.23554

Mummendey, H. D. (2006). Psychologie des „Selbst ": Theorien, Methoden und Ergebnisse der Selbstkonzeptforschung. Göttingen, Bern, Wien, Toronto, Seattle. In: Oxford, Prag: Hogrefe Verlag GmbH & Co. KG.

Murdaugh, D. L., Cox, J. E., Cook, E. W., 3rd, & Weller, R. E. (2012). fMRI reactivity to high-calorie food pictures predicts short- and long-term outcome in a weight-loss program. *Neuroimage*, *59*(3), 2709-2721. https://doi.org/10.1016/j.neuroimage.2011.10.071

Neseliler, S., Hu, W., Larcher, K., Zacchia, M., Dadar, M., Scala, S. G., . . . Dagher, A. (2019). Neurocognitive and Hormonal Correlates of Voluntary Weight Loss in Humans. *Cell Metab*, *29*(1), 39-49.e34. https://doi.org/10.1016/j.cmet.2018.09.024

Ness, A., Bruce, J., Bruce, A., Aupperle, R., Lepping, R., Martin, L., . . . Savage, C. R. (2014). Pre-surgical cortical activation to food pictures is associated with weight loss following bariatric surgery. *Surg Obes Relat Dis*, *10*(6), 1188-1195. https://doi.org/10.1016/j.soard.2014.06.005

Nock, N. L., Dimitropolous, A., Tkach, J., Frasure, H., & von Gruenigen, V. (2012). Reduction in neural activation to high-calorie food cues in obese endometrial cancer survivors after a behavioral lifestyle intervention: a pilot study. *BMC Neurosci*, *13*, 74. https://doi.org/10.1186/1471-2202-13-74

Noori, H. R., Cosa Linan, A., & Spanagel, R. (2016). Largely overlapping neuronal substrates of reactivity to drug, gambling, food and sexual cues: A comprehensive meta-analysis. *Eur Neuropsychopharmacol*, *26*(9), 1419-1430. https://doi.org/10.1016/j.euroneuro.2016.06.013

Noori, H. R., Spanagel, R., & Hansson, A. C. (2012). Neurocircuitry for modeling drug effects [Research Support, Non-U.S. Gov't

Review]. *Addict Biol*, *17*(5), 827-864. https://doi.org/10.1111/j.1369-1600.2012.00485.x

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.

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain--a meta-analysis of imaging studies on the self. *Neuroimage*, *31*(1), 440-457. https://doi.org/10.1016/j.neuroimage.2005.12.002

O'Herron, P., Chhatbar, P. Y., Levy, M., Shen, Z., Schramm, A. E., Lu, Z., & Kara, P. (2016). Neural correlates of single-vessel haemodynamic responses in vivo. *Nature*, *534*(7607), 378-382. https://doi.org/10.1038/nature17965

Ochner, C. N., Kwok, Y., Conceição, E., Pantazatos, S. P., Puma, L. M., Carnell, S., . . . Geliebter, A. (2011). Selective reduction in neural responses to high calorie

foods following gastric bypass surgery. *Ann Surg*, *253*(3), 502-507. https://doi.org/10.1097/SLA.0b013e318203a289

Ochner, C. N., Laferrère, B., Afifi, L., Atalayer, D., Geliebter, A., & Teixeira, J. (2012). Neural responsivity to food cues in fasted and fed states pre and post gastric bypass surgery. *Neurosci Res*, *74*(2), 138-143. https://doi.org/10.1016/j.neures.2012.08.002

Ochner, C. N., Stice, E., Hutchins, E., Afifi, L., Geliebter, A., Hirsch, J., & Teixeira, J. (2012). Relation between changes in neural responsivity and reductions in desire to eat high-calorie foods following gastric bypass surgery. *Neuroscience*, *209*, 128-135. https://doi.org/10.1016/j.neuroscience.2012.02.030

Ombao, H., Lindquist, M., Thompson, W., & Aston, J. (2016). *Handbook of Neuroimaging Data Analysis*. CRC Press. https://books.google.de/books?id=khcNDgAAQBAJ

Pelchat, M. L., Johnson, A., Chan, R., Valdez, J., & Ragland, J. D. (2004). Images of desire: food-craving activation during fMRI. *Neuroimage*, *23*(4), 1486-1493.

Peter, J. P., Churchill Jr, G. A., & Brown, T. J. (1993). Caution in the use of difference scores in consumer research. *Journal of consumer research*, *19*(4), 655-662.

Rammsayer, T., & Weber, H. (2016). *Differentielle Psychologie–Persönlichkeitstheorien* (Vol. 1). Hogrefe Verlag.

Rapuano, K. M., Zieselman, A. L., Kelley, W. M., Sargent, J. D., Heatherton, T. F., & Gilbert-Diamond, D. (2017). Genetic risk for obesity predicts nucleus accumbens size and responsivity to real-world food cues. *Proceedings of the National Academy of Sciences*, *114*(1), 160-165.

Rath, J., Wurnig, M., Fischmeister, F., Klinger, N., Höllinger, I., Geißler, A., . . . Beisteiner, R. (2016). Between- and within-site variability of fMRI localizations. *Hum Brain Mapp*, *37*(6), 2151-2160. https://doi.org/10.1002/hbm.23162

Regier, D. A., Narrow, W. E., Clarke, D. E., Kraemer, H. C., Kuramoto, S. J., Kuhl, E. A., & Kupfer, D. J. (2013). DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. *Am J Psychiatry*, *170*(1), 59-70. https://doi.org/10.1176/appi.ajp.2012.12070999

Reinhard, I., Lemenager, T., Fauth-Buhler, M., Hermann, D., Hoffmann, S., Heinz, A., . . . Vollstadt-Klein, S. (2015). A comparison of region-of-interest measures for extracting whole brain data using survival analysis in alcoholism as an example [Clinical Trial

Research Support, Non-U.S. Gov't]. *J Neurosci Methods*, *242*, 58-64. https://doi.org/10.1016/j.jneumeth.2015.01.001

Rindermann, H. (2009). *Emotionale-Kompetenz-Fragebogen*. Hogrefe.

Rombouts, S., Barkhof, F., Hoogenraad, F., Sprenger, M., Valk, J., & Scheltens, P. (1997). Test-retest analysis with functional MR of the activated area in the human visual cortex. *American journal of neuroradiology*, *18*(7), 1317-1322.

Rombouts, S. A., Barkhof, F., Hoogenraad, F. G., Sprenger, M., & Scheltens, P. (1998). Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. *Magnetic resonance imaging*, *16*(2), 105-113.

Rosenberg, M. J. (1965). *Society and the adolescent self-image*. Princeton University Press.

Schacht, J. P., Anton, R. F., Randall, P. K., Li, X., Henderson, S., & Myrick, H. (2011). Stability of fMRI striatal response to alcohol cues: a hierarchical linear modeling approach. *Neuroimage*, *56*(1), 61-68. https://doi.org/10.1016/j.neuroimage.2011.02.004

Schacht, J. P., Randall, P. K., Latham, P. K., Voronin, K. E., Book, S. W., Myrick, H., & Anton, R. F. (2017). Predictors of Naltrexone Response in a Randomized Trial: Reward-Related Brain Activation, OPRM1 Genotype, and Smoking Status. *Neuropsychopharmacology*, *42*(13), 2640-2653. https://doi.org/10.1038/npp.2017.74

Schneider, F. (2013). *Funktionelle MRT in Psychiatrie und Neurologie*.

Schulte, E. M., Yokum, S., Jahn, A., & Gearhardt, A. N. (2019). Food cue reactivity in food addiction: A functional magnetic resonance imaging study. *Physiol Behav*, *208*, 112574. https://doi.org/10.1016/j.physbeh.2019.112574

Seo, S., Mohr, J., Beck, A., Wüstenberg, T., Heinz, A., & Obermayer, K. (2015). Predicting the future relapse of alcohol-dependent patients from structural and functional brain images. *Addict Biol*, *20*(6), 1042-1055. https://doi.org/10.1111/adb.12302

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*, *86*(2), 420-428. https://doi.org/10.1037//0033-2909.86.2.420

Smeets, P. A. M., Dagher, A., Hare, T. A., Kullmann, S., van der Laan, L. N., Poldrack, R. A., . . . Veldhuizen, M. G. (2019). Good practice in food-related neuroimaging. *Am J Clin Nutr*, *109*(3), 491-503. https://doi.org/10.1093/ajcn/nqy344

Sobell, L. C., Brown, J., Leo, G. I., & Sobell, M. B. (1996). The reliability of the Alcohol Timeline Followback when administered by telephone and by computer [Comparative Study]. *Drug Alcohol Depend*, *42*(1), 49-54. http://www.ncbi.nlm.nih.gov/pubmed/8889403

Spielberger, C. (1983). *Manual for the State-Trait Anxiety Inventory*. Consulting Psychologists Press.

Sprooten, E., Rasgon, A., Goodman, M., Carlin, A., Leibu, E., Lee, W. H., & Frangou, S. (2017). Addressing reverse inference in psychiatric neuroimaging: Meta-analyses of task-related brain activation in common mental disorders. *Hum Brain Mapp*, *38*(4), 1846-1864. https://doi.org/10.1002/hbm.23486

Swarm Jr, W. B. (1983). Self-verification: Bringing social reality into harmony with the self. *Social psychological perspectives on the self*, *2*, 33-66.

Szabo-Reed, A. N., Martin, L. E., Hu, J., Yeh, H. W., Powell, J., Lepping, R. J., . . . Savage, C. R. (2020). Modeling interactions between brain function, diet adherence behaviors, and weight loss success. *Obes Sci Pract*, *6*(3), 282-292. https://doi.org/10.1002/osp4.403

Tregellas, J. R., Wylie, K. P., Rojas, D. C., Tanabe, J., Martin, J., Kronberg, E., . . . Cornier, M. A. (2011). Altered default network activity in obesity. *Obesity (Silver Spring)*, *19*(12), 2316-2321. https://doi.org/10.1038/oby.2011.119

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., . . . Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, *15*(1), 273-289. https://doi.org/10.1006/nimg.2001.0978

Van Der Laan, L. N., & Smeets, P. A. (2015). You are what you eat: a neuroscience perspective on consumers' personality characteristics as determinants of eating behavior. *Current Opinion in Food Science*, *3*, 11-18.

Vazquez, A. L., & Noll, D. C. (1998). Nonlinear aspects of the BOLD response in functional MRI. *Neuroimage*, *7*(2), 108-118. https://doi.org/10.1006/nimg.1997.0316

Vollstadt-Klein, S., Loeber, S., Kirsch, M., Bach, P., Richter, A., Buhler, M., . . . Kiefer, F. (2011). Effects of cue-exposure treatment on neural cue reactivity in alcohol dependence: a randomized trial. *Biol Psychiatry*, *69*(11), 1060-1066. https://doi.org/10.1016/j.biopsych.2010.12.016

Vollstadt-Klein, S., Loeber, S., Richter, A., Kirsch, M., Bach, P., von der Goltz, C., . . . Kiefer, F. (2012). Validating incentive salience with functional magnetic resonance imaging: association between mesolimbic cue reactivity and attentional bias in alcohol-dependent patients. *Addict Biol*, *17*(4), 807-816. https://doi.org/10.1111/j.1369-1600.2011.00352.x

Voon, V., Grodin, E., Mandali, A., Morris, L., Donamayor, N., Weidacker, K., . . . Momenan, R. (2020). Addictions NeuroImaging Assessment (ANIA): Towards an integrative framework for alcohol use disorder. *Neurosci Biobehav Rev*, *113*, 492-506. https://doi.org/10.1016/j.neubiorev.2020.04.004

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on psychological science*, *4*(3), 274-290.

Wartberg, L., Kriston, L., Kramer, M., Schwedler, A., Lincoln, T. M., & Kammerl, R. (2017). Internet gaming disorder in early adolescence: Associations with parental and adolescent mental health. *European Psychiatry*, *43*, 14-18.

Wartberg, L., Kriston, L., Zieglmeier, M., Lincoln, T., & Kammerl, R. (2019). A longitudinal study on psychosocial causes and consequences of Internet gaming disorder in adolescence. *Psychological medicine*, *49*(2), 287-294.

Weishaupt, D. (2014). *Wie funktioniert MRI?*

Weygandt, M., Spranger, J., Leupelt, V., Maurer, L., Bobbert, T., Mai, K., & Haynes, J. D. (2019). Interactions between neural decision-making circuits predict long-term dietary treatment success in obesity. *Neuroimage*, *184*, 520-534. https://doi.org/10.1016/j.neuroimage.2018.09.058

Whitworth, J. A., Williamson, P. M., Mangos, G., & Kelly, J. J. (2005). Cardiovascular consequences of cortisol excess. *Vasc Health Risk Manag*, *1*(4), 291-299. https://doi.org/10.2147/vhrm.2005.1.4.291

WHO. (2018). *Obesity and overweight*. World Health Organization. Retrieved February from http://www.who.int/mediacentre/factsheets/fs311/en/

Wiemerslage, L., Nilsson, E. K., Solstrand Dahlberg, L., Ence-Eriksson, F., Castillo, S., Larsen, A. L., . . . Bandstein, M. (2016). An obesity-associated risk allele within the FTO gene affects human brain activity for areas important for emotion, impulse control and reward in response to food images. *European Journal of Neuroscience*, *43*(9), 1173-1180.

Wittchen, H. U., Zaudig, M., & Fydrich, T. (1997). *Strukturiertes Klinisches Interview für DSM-IV*. Hogrefe.

Wittchen, H. U., Zaudig, M., & Fydrich, T. (1997). *Strukturiertes Klinisches Interview für DSM-IV (SKID-I und SKID-II) - [The Structured Clinical Interview for DSM-IV (SCID-I and SCID II)]*. Hogrefe.

Wölfling, K., Beutel, M. E., & Müller, K. W. (2012). Construction of a Standardized Clinical Interview to Assess Internet addiction: First Findings Regarding the Usefulness of AICA-C *Addiction Research & Therapy S6:003*. https://doi.org/10.4172/2155-6105.S6-003

World Health Organization. Division of Mental, H. (1992). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines*. World Health Organization.

World Health Organization. Division of Mental, H. (2019). *ICD-11: International classification of diseases (11th revision).* World Health Organization.

Wright, K. (2014). Adjusting effect sizes in light of reliability estimates. Annual Meeting of the Southwest Educational Research Association,

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspect Psychol Sci*, *12*(6), 1100-1122. https://doi.org/10.1177/1745691617693393

Zoon, H. F., de Bruijn, S. E., Smeets, P. A., de Graaf, C., Janssen, I., Schijns, W., . . . Boesveldt, S. (2018). Altered neural responsivity to food cues in relation to food preferences, but not appetite-related hormone concentrations after RYGB-surgery. *Behav Brain Res*, *353*, 194-202.

6 CURRICULUM VITAE

PERSONAL INFORMATION

Name	Patrick Michael Walter Bach
Date of Birth	20.06.1987
Place of Birth	Weinheim
Nationality	German
Father	Prof. Dr. med. Alfons Bach
Mother	Dr. med. Ellinor Ruppenthal-Bach

ACEDEMIC EDUCATION

09.04.2021	Habilitation (Venia legendi *Experimental Psychiatry*, Medical Faculty Mannheim, Heidelberg University)
01.02.2017	Doctoral thesis (MD): *The effects of single nucleotide polymorphisms in glutamatergic neurotransmission genes on neural response to alcohol cues and craving* (summa cum laude)
28. & 29.11.2016	Third medical examination (grade 1.0)
15.10.2015	Secondary medical examination (grade 1.0)
12.09.2012	First medical examination (grade 1.0)
2010 - 2016	Medical studies, Medical Faculty Mannheim, Heidelberg University
14.07.2010	B.Sc. Psychology (grade 1.4), University of Mannheim
14.06.2010	B.Sc. Thesis: *The thermal grill illusion: Role of descending inhibition and implications for thermal allodynia* (grade 1.0)
2007 - 2010	Studies in Psychology at the University of Mannheim

PRIMARY AND SECONDARY EDUCATION

26.06.2007	University-entrance diploma (grade 1.1)
1998-2007	Carl Benz Gymnasium, Ladenburg
1994-1998	Dalberg Grundschule, Ladenburg

7 PUBLICATION LIST

Included Publications

Bach P, Reinhard I, Koopmann A, Bumb JM, Sommer WH, Vollstädt-Klein S, Kiefer F. Test-retest reliability of neural alcohol cue-reactivity: Is there light at the end of the magnetic resonance imaging tube? Addict Biol. 2022 Jan;27(1):e13069. doi: 10.1111/adb.13069. Epub 2021 Jun 15. PMID: 34132011.

Bach P, Grosshans M, Koopmann A, Kienle P, Vassilev G, Otto M, Bumb JM, Kiefer F. Reliability of neural food cue-reactivity in participants with obesity undergoing bariatric surgery: a 26-week longitudinal fMRI study. Eur Arch Psychiatry Clin Neurosci. 2021 Aug;271(5):951-962. doi: 10.1007/s00406-020-01218-8. Epub 2020 Dec 17. PMID: 33331960; PMCID: PMC8236041.

Bach P, Hill H, Reinhard I, Gädeke T, Kiefer F, Leménager T. Reliability of the fMRI-based assessment of self-evaluation in individuals with internet gaming disorder. Eur Arch Psychiatry Clin Neurosci. 2021 Jul 17. doi: 10.1007/s00406-021-01307-2. Epub ahead of print. PMID: 34275007.

Other Publications

Bach P., Schuster R., Koopmann A., Vollstädt-Klein S., Spanagel R., Kiefer F. Plasma calcium concentration during detoxification predicts neural cue-reactivity and craving during early abstinence in alcohol-dependent patients. *Eur Arch Psychiatry Clin Neurosci.* **272**(2):341-348 (2022).

Zimmermann S.Z., Thomas B.C., Krisam J., Limprecht R., Klose C., Stenger M., Pourbaix M., Ries M., Vollstädt-Klein S., Koopmann A., Lenz B., Kiefer F., **Bach P**. ON-ICE trial: Investigation of the combined effects of oxytocin and naltrexone on stress-induced and alcohol cue-induced craving in alcohol use

disorder-Study protocol of a phase II randomised double-blind placebo-controlled parallel-group trial. *BMJ Open* **12**(4):e059672 (2022).

Bach, P., Weil, G., Pompili, E., Hoffmann, S., Hermann, D., Vollstädt-Klein, S., Kiefer, F., Mann, K., Sommer, W.H. FMRI-based prediction of naltrexone response in alcohol use disorder: a replication study. *Eur Arch Psychiatry Clin Neurosci* **271**(5):915-927 (2021).

Bach, P., Grosshans, M., Koopmann, A., Pfeifer, A.M., Vollstadt-Klein, S., Otto, M., Kienle, P., Bumb, J.M., Kiefer, F. Predictors of weight loss in participants with obesity following bariatric surgery - A prospective longitudinal fMRI study. Appetite **163**: 105237 (2021).

Bumb J.M.*, **Bach P.***, Grosshans M., Wagner X., Koopmann A., Vollstädt-Klein S., Schuster R., Wiedemann K., Kiefer F. BDNF influences neural cue-reactivity to food stimuli and food craving in obesity. *Eur Arch Psychiatry Clin Neurosci*. **271**(5):963-974 (2021). * = equally contributing authors

Bach, P., Koopmann, A., Bumb, J.M., Vollstädt-Klein, S., Reinhard, I., Rietschel, M., Witt, S.H., Wiedemann, K., Kiefer, F. Leptin predicts cortical and subcortical gray matter volume recovery in alcohol dependent patients: A longitudinal structural magnetic resonance imaging study. *Hormones and Behavior* **124**: 104749 (2020).

Bach, P., Koopmann, A., Bumb, J.M., Zimmermann, S., Buhler, S., Reinhard, I., Witt, S.H., Rietschel, M., Vollstadt-Klein, S., Kiefer, F. Oxytocin attenuates neural response to emotional faces in social drinkers: an fMRI study. *Eur Arch Psychiatry Clin Neurosci* [Epub ahead of print] (2020).

Bach, P., Reinhard, I., Buhler, S., Vollstadt-Klein, S., Kiefer, F., Koopmann, A. Oxytocin modulates alcohol-cue induced functional connectivity in the nucleus accumbens of social drinkers. *Psychoneuroendocrinology* **109**, 1-5 [Epub ahead of print] (2019).

Bach, P., Weil, G., Pompili, E., Hoffmann, S., Hermann, D., Vollstädt-Klein, S., Mann, K., Perez-Ramirez, U., Moratal, D., Canals, S., Dursun, S.M., Greenshaw, A.J., Kirsch, P., Kiefer, F., Sommer, W.H. Incubation of neural alcohol cue reactivity after withdrawal and its blockade by naltrexone. *Addiction biology* doi:10.1111/adb.12717 [Epub ahead of print] (2019).

Bach, P., Bumb, J. M., Schuster, R., Vollstädt-Klein, S., Reinhard, I., Rietschel, M., Witt, S. H., Wiedemann, K., Kiefer, F. & Koopmann, A. Effects of leptin and ghrelin on neural cue-reactivity in alcohol addiction: Two streams merge to one river? *Psychoneuroendocrinology* **100**, 1-9 [Epub ahead of print] (2019).

Bach, P., Zois, E., Vollstadt-Klein, S., Kirsch, M., Hoffmann, S., Jorde, A., Frank, J., Charlet, K., Treutlein, J., Beck, A., Heinz, A., Walter, H., Rietschel, M. & Kiefer, F. Association of the alcohol dehydrogenase gene polymorphism rs1789891 with gray matter brain volume, alcohol consumption, alcohol craving and relapse risk. *Addiction biology* **24**, 110-120 (2019).

Bach, P., Frischknecht, U., Bungert, M., Karl, D., Vollmert, C., Vollstadt-Klein, S., Lis, S., Kiefer, F., Hermann, D. Effects of social exclusion and physical pain in chronic opioid maintenance treatment: fMRI correlates. *European neuropsychopharmacology: the journal of the European College of Neuropsychopharmacology* **29**, 291-305 (2019).

Bach, P., Frischknecht, U., Reinhard, I., Bekier, N., Demirakca, T., Ende, G., Vollstädt-Klein, S., Kiefer, F., Hermann, D. Impaired working memory performance in opioid-dependent patients is related to reduced insula gray matter volume: a voxel-based morphometric study. *European Archives of Psychiatry and Clinical Neuroscience* doi: 10.1016/j.euroneuro.2018.11.1109 [Epub ahead of print] (2019).

Bach, P., Kirsch, M., Hoffmann, S., Jorde, A., Mann, K., Frank, J., Charlet, K., Beck, A., Heinz, A., Walter, H., Rietschel, M., Kiefer, F. & Vollstadt-Klein, S. The effects of single nucleotide polymorphisms in glutamatergic neurotransmission

genes on neural response to alcohol cues and craving. *Addiction biology* **20**, 1022-1032 (2015).

Bach, P., Vollstädt-Klein, S., Kirsch, M., Hoffmann, S., Jorde, A., Frank, J., Charlet, K., Beck, A., Heinz, A., Walter, H., Sommer, W. H., Spanagel, R., Rietschel, M. & Kiefer, F. Increased mesolimbic cue-reactivity in carriers of the mu-opioid-receptor gene OPRM1 A118G polymorphism predicts drinking outcome: a functional imaging study in alcohol dependent subjects. *European neuropsychopharmacology: the journal of the European College of Neuropsychopharmacology* **25**, 1128-1135 (2015).

PUBLICATIONS (CO-AUTHOR)

Ekhtiari H, Zare-Bidoky M, Sangchooli A, Janes AC, Kaufman MJ, Oliver JA, Prisciandaro JJ, Wüstenberg T, Anton RF, **Bach P**, Baldacchino A, Beck A, Bjork JM, Brewer J, Childress AR, Claus ED, Courtney KE, Ebrahimi M, Filbey FM, Ghahremani DG, Azbari PG, Goldstein RZ, Goudriaan AE, Grodin EN, Hamilton JP, Hanlon CA, Hassani-Abharian P, Heinz A, Joseph JE, Kiefer F, Zonoozi AK, Kober H, Kuplicki R, Li Q, London ED, McClernon J, Noori HR, Owens MM, Paulus MP, Perini I, Potenza M, Potvin S, Ray L, Schacht JP, Seo D, Sinha R, Smolka MN, Spanagel R, Steele VR, Stein EA, Steins-Loeber S, Tapert SF, Verdejo-Garcia A, Vollstädt-Klein S, Wetherill RR, Wilson SJ, Witkiewitz K, Yuan K, Zhang X, Zilverstand A. A methodological checklist for fMRI drug cue reactivity studies: development and expert consensus. *Nat Protoc*. **17**(3):567-595. (2022).

Bordier C., Weil G., **Bach P.**, Scuppa G., Nicolini C., Forcellini G., Perez-Ramirez U., Moratal D., Canals S., Hoffmann S., Hermann D., Vollstädt-Klein S., Kiefer F., Kirsch P., Sommer W.H., Bifone A. Increased network centrality of the anterior insula in early abstinence from alcohol. *Addict Biol* **27**(1):e13096. (2022).

Karl D, Bumb JM, **Bach P**, Dinter C, Koopmann A, Hermann D, Mann KF, Kiefer F, Vollstädt-Klein S. Nalmefene attenuates neural alcohol cue-reactivity in the ventral striatum and subjective alcohol craving in patients with alcohol use disorder. *Psychopharmacology (Berl)*. **238**(8):2179-2189 (2021)

Lenz B, Weinland C, **Bach P**, Kiefer F, Grinevich V, Zoicas I, Kornhuber J, Mühle C. Oxytocin blood concentrations in alcohol use disorder: A cross-sectional, longitudinal, and sex-separated study. Eur Neuropsychopharmacol. **51**:55-67 (2021).

Schuster R, Winkler M, Koopmann A, **Bach P**, Hoffmann S, Reinhard I, Spanagel R, Bumb JM, Sommer WH, Kiefer F. Calcium Carbonate Attenuates Withdrawal and Reduces Craving: A Randomized Controlled Trial in Alcohol-Dependent Patients. *Eur Addict Res*. **27**(5):332-340 (2021).

De Santis, S., **Bach, P**., Perez-Cervera, L., Cosa-Linan, A., Weil, G., Vollstadt-Klein, S., Hermann, D., Kiefer, F., Kirsch, P., Ciccocioppo, R., Sommer, W.H., Canals, S. Microstructural White Matter Alterations in Men With Alcohol Use Disorder and Rats With Excessive Alcohol Consumption During Early Abstinence. *JAMA psychiatry* **76**, 749-758 (2019).

Koopmann, A., **Bach, P**., Schuster, R., Bumb, J. M., Vollstädt-Klein, S., Reinhard, I., Rietschel, M., Witt, S. H., Wiedemann, K. & Kiefer, F. Ghrelin modulates mesolimbic reactivity to alcohol cues in alcohol-addicted subjects: a functional imaging study. *Addiction biology* **24,**1066-1076 (2019).

Koopmann, A., Lippmann, K., Schuster, R., Reinhard, I., **Bach, P**., Weil, G., Rietschel, M., Witt, S. H., Wiedemann, K. & Kiefer, F. Drinking water to reduce alcohol craving? A randomized controlled study on the impact of ghrelin in mediating the effects of forced water intake in alcohol addiction. *Psychoneuroendocrinology* **85**, 56-62 (2017).

Zois, E., Vollstadt-Klein, S., Hoffmann, S., Reinhard, I., **Bach, P**., Charlet, K., Beck, A., Treutlein, J., Frank, J., Jorde, A., Kirsch, M., Degenhardt, F., Walter, H.,

Heinz, A. & Kiefer, F. GATA4 variant interaction with brain limbic structure and relapse risk: A voxel-based morphometry study. *European neuropsychopharmacology : the journal of the European College of Neuropsychopharmacology* **26,** 1431-1437 (2016).

Kiefer, F., Kirsch, M., **Bach, P**., Hoffmann, S., Reinhard, I., Jorde, A., von der Goltz, C., Spanagel, R., Mann, K., Loeber, S. & Vollstadt-Klein, S. Effects of D-cycloserine on extinction of mesolimbic cue reactivity in alcoholism: a randomized placebo-controlled trial. *Psychopharmacology* **232,** 2353-2362 (2015).

Jorde, A., **Bach, P**., Witt, S. H., Becker, K., Reinhard, I., Vollstadt-Klein, S., Kirsch, M., Hermann, D., Charlet, K., Beck, A., Wimmer, L., Frank, J., Treutlein, J., Spanagel, R., Mann, K., Walter, H., Heinz, A., Rietschel, M. & Kiefer, F. Genetic variation in the atrial natriuretic peptide transcription factor GATA4 modulates amygdala responsiveness in alcohol dependence. *Biological psychiatry* **75,** 790-79 (2014).

Grosshans, M., Vollmert, C., Vollstadt-Klein, S., Tost, H., Leber, S., **Bach, P**., Buhler, M., von der Goltz, C., Mutschler, J., Loeber, S., Hermann, D., Wiedemann, K., Meyer-Lindenberg, A. & Kiefer, F. Association of leptin with food cue-induced activation in human reward pathways. *Archives of general psychiatry* **69,** 529-537 (2012).

Vollstadt-Klein, S., Loeber, S., Richter, A., Kirsch, M., **Bach, P**., von der Goltz, C., Hermann, D., Mann, K. & Kiefer, F. Validating incentive salience with functional magnetic resonance imaging: association between mesolimbic cue reactivity and attentional bias in alcohol-dependent patients. *Addiction biology* **17,** 807-816 (2012).

Vollstadt-Klein, S., Loeber, S., Kirsch, M., **Bach, P**., Richter, A., Buhler, M., von der Goltz, C., Hermann, D., Mann, K. & Kiefer, F. Effects of cue-exposure

treatment on neural cue reactivity in alcohol dependence: a randomized trial. *Biological psychiatry* **69,** 1060-1066 (2011).

8 ACKNOWLEDGMENTS

Mein außerordentlicher Dank gilt **Herrn Prof. Dr. med. Falk Kiefer,** Direktor der Klinik für Abhängiges Verhalten und Suchtmedizin am Zentralinstitut für Seelische Gesundheit Mannheim und Lehrstuhlinhaber des Lehrstuhls für Suchtforschung der Universität Heidelberg, für die Überlassung des Themas und der Arbeitsmittel, die mentorielle und fachliche Betreuung meiner Arbeit und die Unterstützung bei der Durchführung eigener Studien. Weiterer Dank gilt **Frau apl. Prof. Dr. sc. hum. Sabine Vollstädt-Klein,** Arbeitsgruppenleiterin und Mitarbeiterin der Klinik für abhängiges Verhalten und Suchtmedizin am Zentralinstitut für Seelische Gesundheit Mannheim, für Ihre mentorielle Supervision über viele Jahre und den regen Austausch bei der Konzeption und der kritischen Revision der Bildgebungsanalysen. Besonderer Dank gilt auch **Herrn apl. Prof. Dr. med. Derik Hermann,** ehemaliger leitender Oberarzt der Klinik für Abhängiges Verhalten und Suchtmedizin am Zentralinstitut für Seelische Gesundheit Mannheim, für seine fachliche und wissenschaftliche Unterstützung meiner Arbeiten. Mein Dank gilt ferner **Herrn apl. Prof. Dr. med. Wolfgang Sommer,** stellevertretender Leiter des Instituts für Psychopharmakologie am Zentralinstitut für Seelische Gesundheit Mannheim, für die außerordentliche Unterstützung und Kooperation im Rahmen mehrerer der vorgelegten Arbeiten sowie den anregenden wissenschaftlichen Diskurs und die kritische Revision meiner Arbeiten. Mein Dank gilt auch **Frau PD Dr. med. Anne Koopmann** und **Herrn PD Dr. med. Malte Bumb,** für ihre fachliche Unterstützung, für den regen wissenschaftlichen Austausch und die kritische Revision meiner Arbeiten. Ferner danke ich **Herrn Dr. sc. hum. Ulrich Frischknecht,** für den wissenschaftlichen Austausch und die Unterstützung im Rahmen mehrerer Projekte. Mein besonderer Dank gilt **Frau Iris Reinhard,** Mitarbeiterin der Abteilung für Biostatistik am Zentralinstitut für Seelische Gesundheit Mannheim, für den wissenschaftlichen Austausch und die Kooperation bei den statistischen Auswertungen. Darüber hinaus danke ich **Frau Dr. sc. hum. Martina Kirsch**, **Frau Sabine Hoffmann, Herrn Enrico Pompili, Frau Sina Zimmermann, Frau Carola Hallmann, Frau Anne Jorde, Frau Dr. med. Rilana Schuster**, **Fr. Sina Bühler** und **Herrn Georg Weil** für die

Unterstützung bei der Probandenrekrutierung und -untersuchung sowie der Unterstützung beim Datenmanagement. Weiterhin danke ich ganz herzlich **Frau Birgit Freudenberger, Frau Birgit Hrinkow** und **Frau Heike Grün** für ihre Unterstützung bei administrativen Angelegenheiten.

Mein besonderer Dank gilt auch meiner Familie, meiner **Frau Marisa Bach** und meinem **Sohn Julian Bach** und meiner **Tochter Fiona Bach** sowie meiner Mutter, Tante und meinen Schwiegereltern für Ihre geduldige und verständnisvolle Unterstützung.

Mein außerordentlicher Dank gilt nicht zuletzt **allen Patientinnen und Patienten** sowie **allen Probandinnen und Probanden,** die bereit waren, an den Studien teilzunehmen.

9 APPENDIX

Statement of own contribution to the Empirical Studies and Publications

Work steps	Study 1 / Publication 1	Study 2 / Publication 2	Study 3 / Publication 3
Concept	Partly; for the reliability analysis part completely	Partly; for the reliability analysis part completely	Partly; for the reliability analysis part completely
Literature review	Completely	Completely	Completely
Data collection	Partly	Partly	Partly
Data analysis	Completely	Completely	Completely
Interpretation of results	Predominantly	Predominantly	Predominantly
Preparation of manuscript	Predominantly	Predominantly	Predominantly
Revision of manuscript	Predominantly	Predominantly	Predominantly