NeuroImage 195 (2019) 174–189

Contents lists available at ScienceDirect

NeuroImage

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

Addressing the reliability fallacy in fMRI: Similar group effects may arise from unreliable individual effects

Juliane H. Fröhner [a,*], Vanessa Teckentrup [b], Michael N. Smolka [a], Nils B. Kroemer [a,b,**]

[a] Department of Psychiatry and Psychotherapy, Technische Universität Dresden, Dresden, Germany
[b] Department of Psychiatry and Psychotherapy, University of Tübingen, Tübingen, Germany

ARTICLE INFO

Keywords:
Reward
fMRI
Biomarker
Variability
Individual reliability
Fmreli toolbox

ABSTRACT

To cast valid predictions of future behavior or diagnose disorders, the reliable measurement of a "biomarker" such as the brain activation to prospective reward is a prerequisite. Surprisingly, only a small fraction of functional magnetic resonance imaging (fMRI) studies report or cite the reliability of brain activation maps involved in group analyses. Here, using simulations and exemplary longitudinal data of 126 healthy adolescents performing an intertemporal choice task, we demonstrate that reproducing a group activation map over time is not a sufficient indication of reliable measurements at the individual level. Instead, selecting regions based on significant main effects at the group level may yield estimates that fail to reliably capture individual variance in the subjective evaluation of an offer. Collectively, our results call for more attention on the reliability of supposed biomarkers at the level of the individual. Thus, caution is warranted in employing brain activation patterns prematurely for clinical applications such as diagnosis or tailored interventions before their reliability has been conclusively established by large-scale studies. To facilitate assessing and reporting of the reliability of fMRI contrasts in future studies, we provide a toolbox that incorporates common measures of global and local reliability.

1. Introduction

Since the early 1990s, researchers have used functional magnetic resonance imaging (fMRI) to characterize general aspects of brain function which are immutable (or "fixed") within a population. Hence, many paradigms were optimized for low between-subject variability (Hedge et al., 2018) typically leading to strong main effects in analyses at the group level. However, the advent of the Research Domain Criteria (RDoC) has led to a surge of interest in individual "biomarkers" for mental disorders (Insel et al., 2010). Nevertheless, the investigation of *intra-individual* variability and stability is still a relatively young, but quickly growing field in fMRI research (Dubois and Adolphs, 2016; Garrett et al., 2013; Kroemer et al., 2016; Van Horn et al., 2008; Vetter et al., 2017). One of the key challenges is to identify an appropriate mapping between individual brain activation and behavior (Finn et al., 2017). A prerequisite for this endeavor is to formally establish that an alleged biomarker, supposed to capture individual neurobiological characteristics of brain function (Insel et al., 2010), can indeed be reliably measured to predict future behavior.

To define such basic statistical requirements for a candidate biomarker, key benchmarks for reliability have been previously established in individual differences research (Dubois and Adolphs, 2016). Reliability reflects the "trustworthiness" of a measure that is to what extent a measure will give the same result across repeated measurements. In contrast, replicability or reproducibility reflect whether a measure will give the same result across different samples or software packages. Reliability is a prerequisite for measurements to be ultimately valid, but it is not well known what the reliability of fMRI brain activation is, even for popular paradigms in the literature (Bennett and Miller, 2010; Vul et al., 2009). Given the variety of study designs, it is pivotal to evaluate and report reliability for each scenario. Reliability is critically important when the scientific objective is to predict or classify as it is often the case in longitudinal or clinical studies (Dubois and Adolphs, 2016). Already in the beginning of the last century, Spearman (1910) pointed out that it is harder to distinguish between persons by a less reliable measure, making it harder to detect associations with other constructs as a result. This

* Corresponding author. Section of Systems Neuroscience, Department of Psychiatry and Psychotherapy, Technische Universität Dresden, Würzburger Str. 35, 01187, Dresden, Germany.
** Corresponding author. Section of Translational Psychiatry, Department of Psychiatry and Psychotherapy, University of Tübingen, Calwerstr. 14, 72076, Tübingen, Germany. Tel.: +49 70712982021; fax: +49 7071295901.
E-mail addresses: Juliane.Froehner@tu-dresden.de (J.H. Fröhner), nils.kroemer@uni-tuebingen.de (N.B. Kroemer).

https://doi.org/10.1016/j.neuroimage.2019.03.053
Received 10 July 2018; Received in revised form 22 March 2019; Accepted 25 March 2019
Available online 28 March 2019
1053-8119/

raises the question to what extent fMRI brain responses could be used to predict treatment outcomes when they are not reliably measured within patients in the first place (Nord et al., 2017). Arguably, not every biomarker must be stable over time to be of clinical use, for example if it reflects an acute state of a disorder. Nevertheless, reliability is mandatory for any risk factor that confers liability and is intended to predict the onset, incidence, and etiology of a disorder.

Psychological measures are commonly regarded as reliable when their reliability exceeds 0.8 (Cicchetti and Sparrow, 1981; Cicchetti, 2001), which is hardly achieved (Hedge et al., 2018). Illustratively, Hedge et al. (2018) have highlighted the antagonism between maximizing robust group-level effects on the one hand ("fixed effect") and reliably detecting individual differences on the other hand ("random effect"). Classical experimental research aims to minimize inter-individual variability by identifying robust effects at the group level and, ideally, in every individual. In contrast, individualized prediction is critically dependent on reproducible differences between individuals, which are captured by random effects in statistical models. Consequently, there is a trade-off between optimizing within-subject or between-subject effects, because they represent independent sources of variance and count as error in the analysis of the other (Yarkoni and Braver, 2010). In terms of reliability, this trade-off can result in two different scenarios. On the one hand, reducing measurement error optimizes main effects by increasing both between- and within-subject reliability. On the other hand, simply optimizing main effects by reducing variance between participants comes at a cost of lower within-subject reliability because the measure does not adequately capture individual differences to reliably discern participants.

With respect to fMRI, generalizable effects were of main interest for a long time. Therefore, researchers focused initially on the reliability of task-based fMRI at the group level (Aron et al., 2006; Fliessbach et al., 2010; Gee et al., 2015). Still, Paul et al. demonstrated recently that even large fMRI studies (i.e. $N = 100$) do not produce group results with good reliability (Paul et al., 2017). Furthermore, previous studies showed that group-level stability is not indicative of individual stability (Raemaekers et al., 2007; van den Bulk et al., 2013; Vetter et al., 2015), while research focusing on within-subject reliability over time has produced mixed results. For example, using resting-state fMRI it has been shown that reliability generally increases with increasing scan duration (Mueller et al., 2015). Good reliability was shown during performance monitoring in adolescents and adults (Koolschijn et al., 2011). Moreover, Plichta et al. (2012) reported differential within-subject reliability for three tasks with similarly high group-level reliability. For two tasks (motivational and cognitive), within-subject reliability was fair to good, whereas reliability was low for the emotional task (Plichta et al., 2012). This is in line with the low reliability reported in three different emotional face processing tasks (Nord et al., 2017). Recently, our group analyzed reliability for three different tasks including a subset of the data reported here (Vetter et al., 2017). They showed good reliability in an emotional attention and an intertemporal choice task, but only fair reliability for a cognitive control task (Vetter et al., 2017). However, their analysis focused on conditions contrasted against baseline and not on parametric or difference contrasts, which are commonly used (Bickel et al., 2009; Hare et al., 2009; Kable and Glimcher, 2007; Wittmann et al., 2010) and thus need to be evaluated regarding their reliability as well. In line with this concern, within-session reliability was recently shown to be high for BOLD responses to faces and to shapes, but alarmingly low for the contrast faces vs. shapes (Infantolino et al., 2018). Relatedly, Gorgolewski et al. (2013) also compared four tasks (motor mapping, covert verb generation, overt word repetition, and landmark identification task) in their within-subject reliability to identify factors affecting between-session variance in fMRI. In line with previous results, they showed differential reliability of tasks and showed that reliability depends on the variability of the underlying cognitive processes (Gorgolewski et al., 2013). However, since Gorgolewski et al. (2013) only tested 10 participants, there is still a great demand for extension in larger longitudinal studies since many participants are needed to provide narrow confidence intervals on the expected reliability, for example for prospective power analyses in clinical trials.

To sum it up, even though the reliability of data is a prerequisite in studying individual differences and fMRI brain activation is increasingly applied to his end, little is known about individual fMRI reliability to date. A straightforward answer to the simple question on the reliability of a given paradigm is further complicated by considerable variability in the existing literature in terms of analysis level (group vs. individual) and reliability measures employed (local measures: e.g., intraclass correlation coefficients (ICC) vs. global measures: e.g., overlap coefficients). Moreover, cross-sectional reliability has received surprisingly little attention so far since low longitudinal reliability might arise from different sources of error (e.g., state effects). Here, we are providing analyses of cross-sectional and longitudinal reliability in simulated data and in a sample of 126 adolescents investigated during an intertemporal choice task at the age of 14, 16 and 18. To facilitate the future comprehensive assessment of fMRI reliability, we introduce the collection of the calculated global and local measures, bundled in the MATLAB toolbox *fmreli* (https://github.com/nkroemer/reliability).

2. Materials and methods

2.1. Longitudinal data case: delay discounting

To illustrate the assessment of reliability of fMRI contrast maps, we used a previously reported longitudinal study. Briefly, adolescents were repeatedly investigated at the ages of 14, 16 and 18, mainly to examine the influence of substance consumption on brain development (for details see Jurk et al., 2018; Ripke et al., 2012, 2014; Rodehacke et al., 2014; Vetter et al., 2015). Recently, we investigated the reliability of selected fMRI contrasts for the first two acquisition waves (Vetter et al., 2017). Vetter et al. (2017) looked at three different tasks using the between-session intraclass correlation coefficient focusing on simple contrasts against baseline. Here, the aim is to provide a substantially extended analysis for various within- and between-session measures of group-level and individual stability using an exemplary paradigm, which is why we focus on the intertemporal choice paradigm across all three acquisition waves.

2.1.1. Participants

The presented data originate from the project "the adolescent brain". To prospectively study brain function and substance use, participants were recruited at the age of 14 years and re-invited at the ages of 16 and 18. During the three acquisition waves, participants underwent an extensive assessment, including fMRI sessions and an intertemporal choice task (Ripke et al., 2012, 2014). Initially, 250 adolescents participated in the study. In total, 151 of them completed the task during every acquisition wave. For the reliability analysis, we excluded 8 participants due to a diagnosis of a mental disorder because the onset might distort individual reliability of candidate biomarkers. Additionally, 17 participants were excluded because they had more than 10% invalid trials in at least one session. In line with previous reports, invalid trials were defined as missing or implausible responses such as deciding for a reward with a subjective value lower than half of the alternative reward (see Ripke et al., 2012). This criterion was imposed to exclude all participants who were not sufficiently attentive or internally consistent in their decisions. Thereby, 126 (60 female) individuals remained for the reliability analysis. This comparably large sample is important for the precision of reliability estimates as demonstrated by a resampling procedure (Figure S.1).

2.1.2. Paradigm

During the intertemporal choice task, participants choose between a smaller immediate amount of money and a larger delayed amount. Participants completed 90 trials resulting in a task duration of 25 min (Fig. 1).

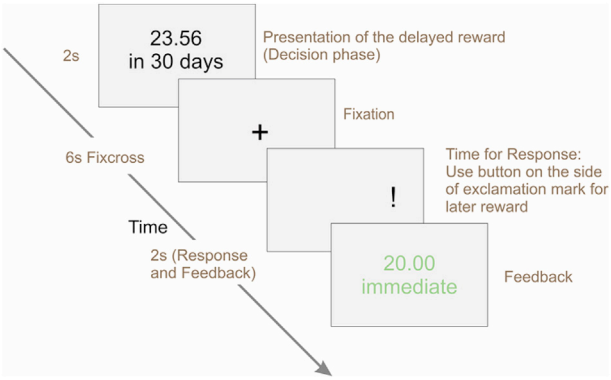

Fig. 1. Time course of one trial in the intertemporal choice task (Ripke et al., 2012). Participants have to choose between a delayed amount of money and a fixed immediate amount (here: 20.00€). Delays ranged from 10 to 180 days and delayed amounts were adapted to each one's discounting rate (for further details, see Supporting Information). After the presentation of the delayed rewards for 2 s, participants have 6 s to decide. Then, an exclamation mark indicates which button (left or right) to press to select the delayed reward. Each trial was followed by an inter-trial interval where a fixation cross was shown on screen (M = 7 s, range [6–8]).

To balance choices for immediate or delayed offers, the offers during the fMRI session were adapted to the individual discounting rate, *k*, determined during a training session at age 14. The temporal discounting rate, *k*, governs the subjective assignment of value, *V*, to a monetary amount *A* when it is delivered after delay *D*:

$$V = \frac{A}{1 + (k*D)} \quad (1)$$

During quality control, we identified that the discount rate was not identical in at least one session for five participants, likely because the discount rate was entered manually before each session. These small discrepancies in the discounting rate led to slightly different offers presented in the task which might influence reliability. However, the effects on the adaptation of the task were very minor and results did not change after excluding these participants. The reason for the robustness of the analyses is that even large changes in the assumed discount rate lead to highly correlated parametric regressors (Figure S.2; for a discussion, see Wilson & Niv, 2015). Thus, these participants were retained for the current analyses.

2.1.3. Behavioral analysis of discounting rate

Reliability of value tracking in the brain might be limited by the reliability of the corresponding behavior. Hence, we assessed the reliability of discounting behavior analogous to the fMRI data. First, we calculated the reliability using Pearson's correlations and consistency ICCs within (split half) and between sessions. Second, we calculated a repeated-measurements analysis of variance (ANOVA) to check whether the discounting rate changes over time. In addition, we simulated choices of participants with correlated discounting rates (N = 126; ICC = 0.5; drawn from a Gaussian distribution with mean $\ln(k) = -4$ and SD = 3). This simulation demonstrated that a moderate reliability of *k* does not necessarily limit the reliability of the subjective-value contrast as the commonly observed small changes in k only lead to minor changes in estimated subjective value (Figure S.2, for details see behavioral results).

2.1.4. fMRI data acquisition and analysis

Functional data was acquired with a 3 T whole-body MR tomograph (Magnetom TRIO, Siemens, Erlangen, Germany) equipped with a 12-channel head coil at the Neuroimaging Centre at TU Dresden. A standard echo planar imaging (EPI) sequence was used for the functional images (repetition time (TR): 2410 ms; echo time (TE): 25 ms; flip angle: 80°; number of slices: 42; slice thickness: 2 mm (1 mm gap); field of view (FoV): $192 \times 192\text{mm}^2$; resampled to voxel size: $3 \times 3\text{x}3\text{mm}^3$). For structural images, a 3D T1-weighted magnetisation-prepared rapid gradient echo (MPRAGE) image was acquired (TR: 1900 ms, TE: 2.26 ms, FOV: 256×256 mm^2, 176 slices, voxel size: $1 \times 1\text{x}1$ mm^3, flip angle: 9°; for details, see Supporting Information).

fMRI data analysis was performed using SPM12 (Wellcome Department of Neuroimaging, London, United Kingdom) and MATLAB R2015a (Mathworks, Inc., Sherborn, MA). The preprocessing followed a standard pipeline including slice-time correction, realignment, coregistration to the respective structural image of the participant, normalization to the standard EPI template (Montreal Neurological Institute, MNI) and smoothing with an isotropic Gaussian kernel (8 mm full-width at half-maximum). The first-level regressors included one regressor for the offer onset and the corresponding parametric modulator. The parameter reflects the subjective value of the presented offer, which was calculated via Equation (1) using the *k* determined at age 14. We used the same discount rate for all three waves because it improves comparison of brain response reliability. Accordingly, we found that there was no significant change overall (see behavioral results). Notably, analogous to reinforcement learning models, most changes in *k* within the observed range only have negligible effects on the estimated subjective value and its corresponding parametric regressors would be highly correlated as a result (Wilson & Niv, 2015). To demonstrate this for our task, we simulated a grid of 126 participants with 3 correlated *k*s (see behavioral results and Figure S.2).

At the end of each trial, an exclamation mark appeared at one side of the screen, indicating where participants had to press to select the presented (delayed) offer. To separate the corresponding motor responses, we included two regressors representing the onsets of button presses with the left and right hand, respectively. We included the six realignment parameters as nuisance regressors. In line with previous research, we focused primarily on the offer and the subjective-value contrasts. Nevertheless, we also report the reliability of the motor contrasts (Figure S.10), which have been previously shown to have good retest reliability (Havel et al., 2006; Loubinoux et al., 2001; Marshall et al., 2004; Waldvogel et al., 2000).

2.2. Simulation

Initially, we validated the toolbox using simulated data that originated from the first wave. Therefore, we used the contrast maps for participants at the age of 14 years and simulated longitudinal changes in contrast maps at the age of 16 and 18, respectively. We simulated different levels of intra-individual stability of the primary contrast of interest subjective value over time (M parameter value = 0.008 and SD = 0.02) using a known range of random Gaussian noise (sigma: 0.01–0.04) and compared similarity matrices for the simulated and the experimental data.

2.3. Analysis workflow for reliability estimation

To illustrate the workflow in the assessment of reliability at different levels, we will describe the cross-sectional and longitudinal analyses of reliability via various measures implemented in the toolbox fmreli (Fig. 2). fmreli is an open-access toolbox available via GitHub (https://github.com/nkroemer/reliability). It has a MATLAB-based graphical user interface (GUI). The toolbox requires SPM12 and the "Tools for NIfTI and ANALYZE image" (https://de.mathworks.com/matlabcentral/fileexchange/8797-tools-for-nifti-and-analyze-image).

Broadly, the implemented reliability measures can be separated into two levels. The first level evaluates a pattern across the whole brain or across a region of interest (ROI), for example correlations of activation maps (voxel-wise matrices). We refer to it as global reliability. The second level evaluates voxel-wise responses among participants, for example correlations of each voxel vector between splits or sessions. We refer to it as local reliability. It can be calculated for voxels within one

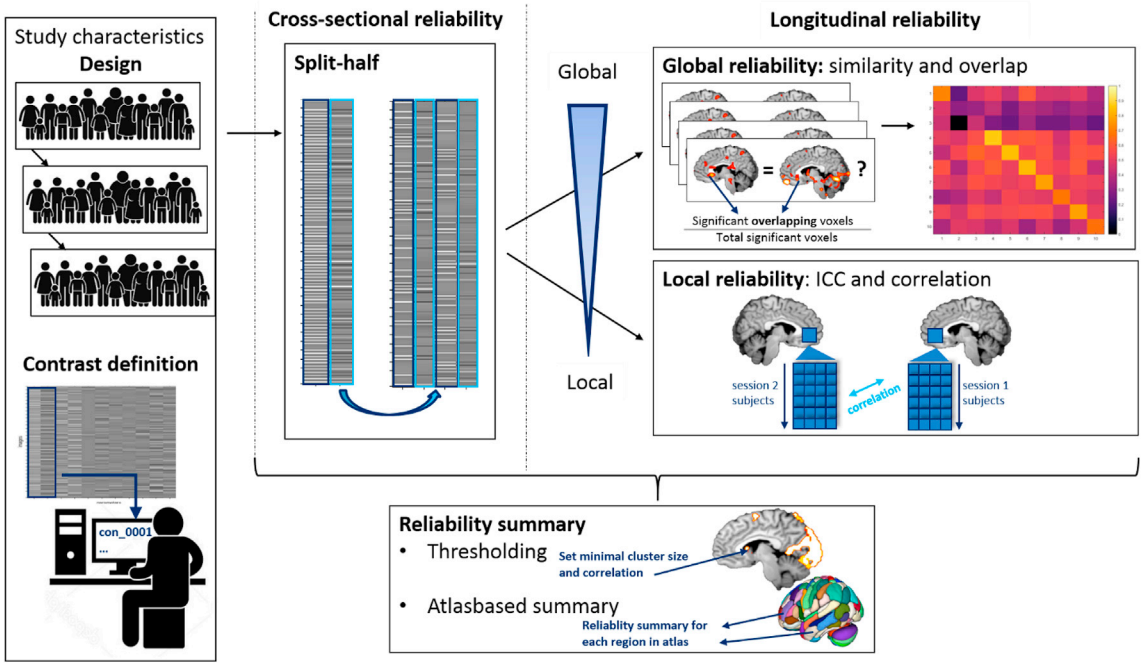

Fig. 2. Schematic summary of the workflow in fmreli. After defining study design and relevant contrasts, regressors can be split for cross-sectional analyses. Analyses are available from a global to a local (voxel-wise) level. Resulting reliability can be summarized based on an anatomical atlas (default: Harvard-Oxford brain atlas plus AAL cerebellum; Whitfield-Gabrieli and Nieto-Castanon, 2012), functional networks (Yeo et al., 2011) or a threshold can be defined to identify regions surpassing a required minimum of reliability.

ROI or for the whole brain. Both levels can be used for the reliability analysis of repeated measurements such as test-retest reliability. We refer to it as longitudinal reliability. In addition, we implemented the option to calculate reliability within one session using a randomized split-half procedure. We refer to it as cross-sectional reliability.

2.3.1. *Preprocessing for within-session analysis: split-half estimation for cross-sectional reliability*

Since longitudinal data is not available in every project and low reliability might also be caused by a substantial delay between repeated measures, we used a split-half estimation procedure to calculate cross-sectional reliability. The defined regressor of interest (i.e., offer onsets) and the corresponding parametric modulator (i.e., subjective value of the offer) were randomly split into two parts while other regressors remained untouched (e.g., motor response and realignment parameters; Fig. 2). To do so, we used the random permutation (randperm) function in MATLAB and divided the re-ordered trials into two parts. Afterwards, first-level statistics were re-estimated. Based on these first-level statistics, we analyzed the reliability of the task-induced brain activation within and between sessions.

To the best of our knowledge, it is not common to assess split-half reliability of fMRI data yet. However, split-half reliability has been evaluated before in other brain imaging modalities such as electroencephalography (EEG) and magnetoencephalography (MEG (Groppe et al., 2009);). Recently, Infantolino et al. (2018) investigated split-half reliability in amygdala signal during an emotional face-matching task. They compared two blocks within one fMRI session, but did not report longitudinal data (Infantolino et al., 2018). There have also been efforts in functional connectivity research to adapt existing reliability measures to single-session data by partitioning the data (Mejia et al., 2018; Mueller et al., 2015). However, current fMRI software packages do not offer a straightforward way to assess split-half reliability for task designs.

Since any statistical dependency between transitions of the partitions will reduce design orthogonality, we checked whether partitioning the data does introduce a problematic inflation of reliability estimates for null data, which we refer to as intercept bias. If the chances that the next trial belongs to the same or the other partition are equal, the intercept bias introduced by the statistical dependency of the signal will be minimal (see Supporting Information, cf. Mumford et al., 2014). Therefore, we used a random partition of trials. Other ways to partition trials can be implemented via the toolbox. For example, comparing early vs. late blocks of trials also leads to low statistical dependence, but might introduce other temporal confounds such as learning or boredom. In the Supporting Information, we first demonstrate the robustness of the split-half procedure across multiple splits. In short, we repeated the analysis using 4 new random splits. After comparing the deviation from the average subjective value and the average split-half correlation, we conclude that the split-half procedure produces highly comparable results after repetition (Figure S.3 and S.4). Second, we show that partitioning the data that way does not introduce an intercept bias, neither for the intertemporal choice task, nor for other commonly used tasks (Figure S.7).

2.3.2. *Global reliability*

A well-established approach to investigate fMRI reliability is the cluster overlap method. Here, a significance level is initially set to define "activated" voxels. Then, the degree of overlap in significant voxels between two measurements is quantified. We used the Dice and the Jaccard coefficient which are commonly used in the literature. The latter can be easily interpreted as the percentage of overlapping significant voxels within all significant voxels (Jaccard, 1901; Maitra, 2010).

$$\text{Overlap}_{\text{Jaccard}} = \frac{V_{overlap}}{V_1 + V_2 - V_{overlap}} \tag{2}$$

The Dice coefficient, first described by Rombouts et al. (1998), is defined as the number of overlapping voxels ($V_{overlap}$) divided by the average number of significant voxels across sessions (V_1, V_2).

$$\text{Overlap}_{\text{Dice}} = \frac{V_{overlap}}{(V_1 + V_2) \times 0.5} \tag{3}$$

Both coefficients range from no overlap (0) to perfect overlap (1). To

the best of our knowledge, there is no consensus for criteria indicating an "acceptable" overlap. In the review by Bennett and Miller (2010), the Dice coefficients ranged from 0.21 to 0.86 for various time lags between measurements (less than 1 h up to 33 weeks), which may serve as a coarse reference for our results. Since the resulting reliability measures are strongly dependent on the significance threshold and the investigated data level, the toolbox offers the option to define the threshold and to calculate it at the individual or the group level. Due to our specific interest in the difference between individual- and group-level data, we analyzed both levels using the uncorrected p .01 as a rather liberal threshold to limit the initial loss of information due to more conservative thresholding ($p < .001$; Table S.4).

As second global measure of reliability, we calculated the similarity of the fMRI activation maps. This has been previously described as part of the representational similarity analysis (Kriegeskorte et al., 2008). Briefly, similarity between activation patterns has been used to characterize the resemblance of the neural representation of objects and categories (Kriegeskorte et al., 2008). In contrast to overlap coefficients, there is no need to specify a threshold for similarity analyses. Hence, similarity can be regarded as a continuous alternative to overlap coefficients. Likewise, Finn et al. (2015) used this approach to compare similarity of connectivity matrices across sessions to successfully re-identify individuals based on maximum similarity ("fingerprinting"). In a similar manner, we compared brain activation matrices within and between subjects and sessions by vectorizing and correlating them. In other words, this procedure captures the resemblance of two patterns based on the alignment of high versus low brain activation values across the brain. Thus, we gathered information about the global reliability of the activation map. For each comparison in each contrast, we checked whether the within-subject similarity is the highest and therefore enables the re-identification of individuals as suggested by Finn et al. (2015). Thus, when a person's activation map is most similar to his or her activation map derived from another set of data, then we count this as a successful re-identification of that person. We then counted how many participants were correctly re-identified for each comparison. Since the similarities are correlation coefficients, they vary between a perfect inverse relationship ($r = -1$) and a perfect direct relationship ($r = 1$). For statistical analyses, similarities were Fisher z-transformed to avoid restriction of variance for higher correlation values and achieve a normal distribution. To visualize similarities, we used color maps where correlations between participants on x- and y-axes are plotted according to the strength of resemblance. In addition, we calculated empirical cumulative density functions via MATLAB (function: ecdf including lower and upper 95% confidence bounds).

2.3.3. Local reliability: ROI- and voxel-based measures

Besides the global approach, reliability can be estimated locally, that is for each voxel in the brain or in a specific ROI. In practice, the level of the reliability analysis should correspond to the level where hypotheses are tested. Since many researchers are focusing on the average activity in a ROI, we compared mean ROI reliability and reliability averaged across voxels. Both measures were highly correlated ($r = 0.715$), albeit not perfectly. Thus, the voxel-based approach may also be useful for studies focusing on mean ROI activation (Figure S.9). However, the feature to compute the reliability of the mean ROI activation will be added in the next update of the toolbox. In our analysis, we evaluated reliability within and outside the significant group-level main effect regions because we expected that intra-individual stability would be to some extent independent of the magnitude of the group-level effect size. Therefore, we calculated each described measure at the voxel and ROI level. The ROI included ventral striatum and ventromedial prefrontal cortex (vmPFC) extracted based on an independent study using the same intertemporal choice paradigm (main effect of subjective value, $p_{uncorrected} < .001$, $k > 20$; Grosskopf et al., submitted).

One important aspect of reliability is captured by the share of total variance being accounted for by the fact that repeated measures are nested within an individual or "class". As introduced by Shrout and Fleiss (1979), the ICC is the ratio of variance of interest and total variance. There are six forms of ICC differing in the defined variance of interest. We were primarily interested in the variance within participants and thus used two different types of ICCs, absolute (ICC_{abs}) and consistency agreement (ICC_{con}). The ICC_{abs} considers the session mean sum of squares ($MS_{session}$) and therefore is the more conservative approach (Equation (4)), whereas the ICC_{con} is only relating the between-subject ($MS_{between}$) to error variance (MS_{error}; Equation (5)). Here, k represents the number of sessions and n represents the number of subjects.

$$ICC_{abs} = \frac{MS_{between} - MS_{error}}{(MS_{between} + (n-1) \times MS_{error}) + \frac{k}{n} \times (MS_{session} - MS_{error})} \quad (4)$$

$$ICC_{con} = \frac{MS_{between} - MS_{error}}{MS_{between} + (n-1) \times MS_{error}} \quad (5)$$

According to guidelines suggested by Fleiss (1986), ICCs lower than 0.4 reflect poor reliability, ICCs between 0.4 and 0.75 reflect fair (<0.6) to good (>0.6) reliability, and ICCs higher than 0.75 reflect excellent reliability (Cicchetti, 2001). Notably, there is already software available to calculate ICCs (Caceres et al., 2009). Still, the primary goal of our toolbox was to make such crucial calculations and the combination with other measures more readily accessible as we envision making reliability analyses an integral part of most neuroimaging research projects.

In addition to ICCs, Pearson's and Spearman's correlation coefficient were calculated. Pearson's and Spearman's correlation coefficients are well-known measures for the strength of association between two variables. Pearson's correlation coefficient is the covariance of two variables that is two vectors containing individual estimates in one voxel for two sessions, divided by the product of their standard deviations. Spearman's correlation coefficient is defined by the Pearson's correlation of the ranked order of the variables. Thus, Pearson's correlation represents the stability of the values in interval scale whereas Spearman's correlation represents the stability of the rank order of the values (Equation (6)). The choice of the correlation coefficient as reliability marker is dependent on the assumptions and applications. If we expect a linear association, Pearson's correlation is recommended. Moreover, Pearson's r is often requested as input to conduct power analyses for repeated measures designs (e.g., in GPower; Faul et al., 2007). Otherwise, Spearman's correlation might be the coefficient of choice as it is more robust to non-linearity of changes across the range.

$$r(x, y) = \frac{cov(x, y)}{\sigma(x) * \sigma(y)} \quad (6)$$

For the analysis of cross-sectional reliability, we applied the Spearman-Brown correction for split-half reliability (Spearman, 1910, Equation (7)) to calculate r_{adj}, accounting for the underestimation of reliability in r_{unadj} due to the decreased number of items.

$$r_{adj}(x, y) = \frac{2 * r_{unadj}}{1 + r_{unadj}} \quad (7)$$

Here, we consider correlations (in absolute value) up to 0.35 as low, up to 0.67 as moderate and above 0.67 as high (Taylor, 1990). Despite their considerable descriptive value, systematic errors might lead to distortion of the coefficients (Safrit, 1976). Pearson's r is an *inter*class coefficient. It reflects whether one variable can be associated with another variable by means of a linear transformation (McGraw and Wong, 1996). In contrast, the ICC is an *intra*class additivity index. It reflects whether one variable can be equated to another variable that is measured on the same scale (and comes from the same "class") by adding a constant (McGraw and Wong, 1996). Another disadvantage of Pearson's and Spearman's correlation coefficients is that they are not readily useable for more than two runs or sessions. Thus, for most purposes, the ICC appears as more suitable.

2.3.4. Summary measures

To aid the assessment of local differences in reliability, several means of aggregation provide useful insights. After calculating the voxel-wise correlations and ICCs, researchers may identify reliable regions by setting minimum reliability thresholds and a minimum cluster extent to restrict the following analysis to these regions by narrowing down the set of candidate voxels.

In addition, the extraction of ROI-based summary statistics is useful, for example to define if brain responses in a specific ROI are a suitable biomarker. Hence, we implemented the option to use an atlas to summarize the results of the toolbox and defined an anatomical atlas for the paper and as default for the toolbox. This reliability summary included all regions in the atlas provided with the CONN functional connectivity toolbox (Whitfield-Gabrieli and Nieto-Castanon, 2012) consisting of the Harvard-Oxford brain atlas (Desikan et al., 2006; Frazier et al., 2005; Goldstein et al., 2007; Makris et al., 2006) and the Automated Anatomic Labeling atlas (AAL) for the cerebellum (Tzourio-Mazoyer et al., 2002). The atlas-based summary might enable choosing a sufficiently reliable anatomical ROI for future analyses. Moreover, it facilitates the comparison of reliability, for example between cortical and subcortical regions, across paradigms or studies, and even modalities and species. Both, the identification of reliable clusters and the reliability summary for anatomical ROIs are implemented in fmreli.

To visualize the summary, we grouped ROIs according to their maximum overlap with functional networks introduced by Yeo et al. (2011). Furthermore, subcortical limbic regions were assigned to the limbic network and we included the cerebellum as an additional network. Finally, we correlated the split-half reliabilities of the offer and the subjective-value contrasts for each region to test if reliabilities at the ROI level are associated (i.e., statistically dependent).

2.3.5. Statistical dependence of reliability and signal amplitude and variance

Since we reasoned that effects at the group level are not necessarily predictive of the reliability at an individual level, we further assessed to what extent the local reliability is related to the average contrast amplitude ("beta") and its variance across individuals. Therefore, we correlated these group-level summary statistics with the z-transformed split-half reliability (*Spearman's rho*) for each voxel. To emphasize the overall association, we show associations pooled across sessions.

2.3.6. fMRI reliability and behavioral consistency

fMRI signal is intended to measure brain function underlying human behavior. Therefore, the degree of behavioral consistency eliciting a certain signal might also influence its reliability such that the signal and its reliability are higher when participants behave more consistently. To test this possibility, we computed a hierarchical linear model predicting similarity within and between subjects with two measures of consistency: the beta (β) parameter from the hyperbolic model (Ripke et al., 2012) and the Bayesian Information Criterion (BIC) as goodness-of-fit of the hyperbolic model. Likewise, we computed a hierarchical linear model to predict the value-tracking signal in the ventral striatum and the offer signal in the dorsolateral prefrontal cortex (dlPFC) with β and BIC. These regions were selected because they have been repeatedly associated with these cognitive functions (e.g. McClure et al., 2004).

3. Results

To determine the reliability of a commonly employed intertemporal choice task across adolescence, we investigated the longitudinal trajectories of discounting behavior and brain response to delayed monetary offers. We calculated longitudinal reliability at the behavioral and brain level. As a reference reflecting momentary reliability, we further used split-half reliability calculated within sessions at each wave. The rationale behind this two-fold approach was to dissociate potential unreliability due to differential developmental trajectories from cross-sectional reliability of the measurement. We reasoned that longitudinal reliability would not exceed cross-sectional reliability. Furthermore, to verify the scripts and constrain plausible results for the reliability of the longitudinal data, we simulated changes using known inputs of signal relative to varying degrees of noise (i.e., half to double the initial signal-to-noise ratio).

3.1. Behavioral results

To estimate the reliability of discounting behavior over the three scanner sessions, we calculated Pearson's correlations between sessions and the consistency ICC. Between-session correlations ($r_{14\text{-}16} = 0.528$, $p < .001$; $r_{16\text{-}18} = 0.514$, $p < .001$; $r_{14\text{-}18} = 0.303$, $p = .001$) and ICCs were in a moderate range (Fig. 3A; $ICC_{14\text{-}16} = 0.529$, $p < .001$; $ICC_{16\text{-}18} = 0.529$, $p < .001$; $ICC_{14\text{-}18} = 0.512$, $p < .001$). Moreover, the repeated-measurements ANOVA revealed no significant change over time, $F(1.823, 227.868) = 0.816$, $p = .43$.

To assess behavioral within-session reliability, we calculated the discounting rate k using the trials of the respective partition only and then correlated them across partitions within one session. Split-half reliability was high for all sessions ($r_{14} = 0.803$; $r_{16} = 0.828$; $r_{18} = 0.867$, $ps < .001$). The ICCs for the two splits in each wave were comparably high (Fig. 3B; $ICC_{14} = 0.672$, $p < .001$; $ICC_{16} = 0.707$, $p < .001$; $ICC_{18} = 0.764$, $p < .001$).

Next, to check whether moderate behavioral reliability would limit fMRI reliability, we simulated changes in k (drawn from a Gaussian distribution with $\ln(k)$ $M = -4$ and $SD = 3$) as they occurred in our sample and correlated the resulting subjective-value regressors. In our sample, the largest change was 4.93 units and changes in $\ln(k)$ of the magnitude ± 5 preserve correlations $\geq .80$ between different subjective-value regressors. Only if changes approximate ± 7.5 units of $\ln(k)$, we saw that the correlation of the subjective-value regressors hit the level of the ICC of the discount rate (Figure S.2, for details see Supporting Information). Since such drastic changes (i.e., from the lowest percentile of the sample to the highest percentile) were not observed, we concluded that a moderate level of reliability would not impair the tracking of subjective value in the brain *per se*.

3.2. Group main effects for subjective value and offer contrasts

First, we observed a strong main effect of the offer onset event, which was highly congruent across the three sessions, encompassing the occipital and parietal cortex, the thalamus, and the dorsal anterior cingulate cortex (dACC). Second, we observed that the parametric modulator "subjective value" of the offer event was positively associated with activation in ventral striatum (peak: −9, 5, −6; $T_{max} = 11.02$; $p_{FWE\text{-}corrected} < 0.05$), vmPFC (peak: −3, 41, 4; $T_{max} = 9.28$; $p_{FWE\text{-}corrected} < 0.05$) and posterior cingulate cortex (peak: 0, −31, 34; $T_{max} = 7.52$; $p_{FWE\text{-}corrected} < 0.05$). Hence, in line with numerous previous studies (e.g., Kofarnus et al., 2017; McClure et al., 2004; Peters and Buchel, 2009; Ripke et al., 2012), there were robust and seemingly congruent main effects for both contrasts across sessions (Fig. 4).

3.3. Global reliability

To assess the overlap of activation maps across sessions, we calculated the Jaccard and Dice coefficients as indices of reliability at a liberal threshold ($p_{uncorrected} < .01$) both individually (i.e., for each participant) and aggregated at the group level (Table 1). For the offer contrast, the Dice overlap ranged from .91 to .96 at the group level, whereas it ranged from 0.39 to 0.61 for the subjective-value contrast. Critically, the congruency of group main effects was in stark contrast to the unreliability of individual activation maps. In particular, the parametric contrast for subjective value yielded very low coefficients of overlap, suggesting that the individual information contained in the contrasts is not reliable (Fig. 5). Applying a more stringent threshold ($p_{uncorrected} < .001$) also lead to worse overlap coefficients, indicating that these results are not due to a

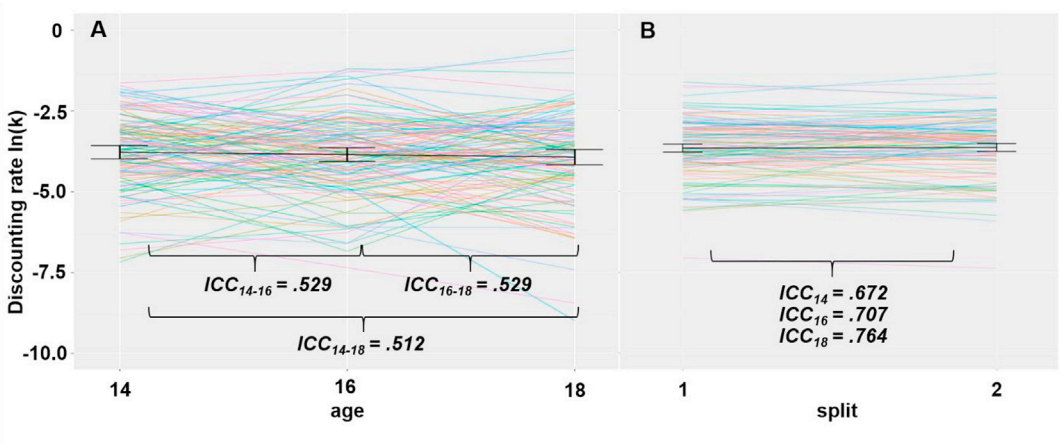

Fig. 3. Individual trajectories (one color per participant) and average discounting, ln(k), with 95% confidence interval (black) for longitudinal data (A) and split-half data pooled for all three sessions (B). While there is a marked variability across individuals, there is no significant change in average discounting behavior and longitudinal reliability is within the moderate range.

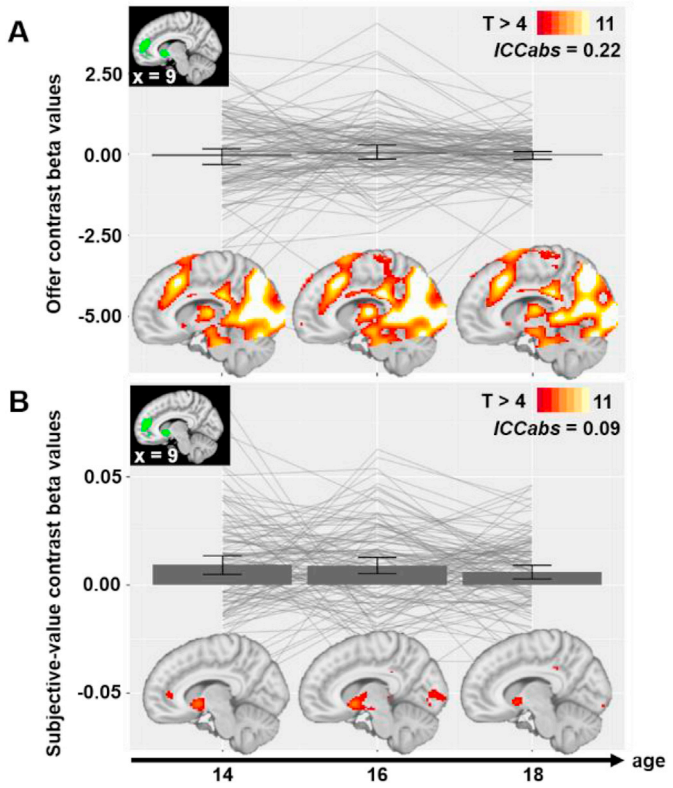

Fig. 4. Individual trajectories of contrast beta values (grey lines) and mean activation (error bars: 95% confidence interval) for a region of interest (ROI, in green) in ventral striatum and vmPFC for the unmodulated (A) offer contrast and (B) the subjective-value contrast. Although the mask was derived to capture regions most strongly linked to subjective-value tracking, individual betas from the unmodulated offer contrast were more reliable in these regions compared to the subjective-value contrast. The ROI was derived from an independent sample using the same paradigm (Grosskopf et al., submitted) to avoid potential over-fitting and selection bias in reliability estimates. Brain images show the respective whole-brain group activation maps.

liberal threshold (for details, see Supporting Information and Table S.4). The notable difference between individual and group-level stability was also visible in the cross-sectional comparisons. In other words, low cross-sectional reliability substantiates that the low longitudinal reliability is

Table 1
Comparison of average individual overlap and group-level overlap for offer and subjective-value contrasts.

Comparison		Group level		Mean individual level	
		Jaccard*	Dice*	Jaccard*	Dice*
		Offer contrast			
Longitudinal	14 to 16	0.85	0.92	0.17	0.26
	14 to 18	0.83	0.91	0.17	0.25
	16 to 18	0.87	0.93	0.18	0.27
Cross-sectional	Split 14	0.89	0.94	0.12	0.17
	Split 16	0.91	0.96	0.13	0.18
	Split 18	0.90	0.95	0.10	0.15
		Subjective-value contrast			
Longitudinal	14 to 16	0.37	0.54	0.02	0.02
	14 to 18	0.30	0.46	0.01	0.02
	16 to 18	0.35	0.52	0.01	0.03
Cross-sectional	Split 14	0.31	0.47	0.01	0.02
	Split 16	0.44	0.61	0.02	0.03
	Split 18	0.56	0.39	0.01	0.02

Note. *Overlap coefficients are calculated for suprathreshold voxels ($p_{uncorrected} < .01$). For overlap coefficients based on a more conservative threshold ($p < .001$), see Table S.4.

not simply explained by low stability of the contrast estimates over time during adolescence. Notably, the low reliability of the brain response associated with parametric value tracking also stands in contrast to the behavioral results, where the ICCs for the discount rate were well within the moderate to high range (see behavioral results).

Due to the marked difference between reliability at the group vs. individual level, we next calculated the similarity (Fisher-z-transformed) between individual activation maps across individuals and sessions. Conceptually, this method enabled us to quantify the resemblance of individual brain activation patterns across individuals and contrasts demonstrating how unique an induced brain response is. Statistically, this method enabled us to quantify within- and between-subject similarity of task-evoked brain activation without the necessity to define an arbitrary statistical threshold. Again, we observed higher similarity within subjects ($.60 \leq r_{average} \leq 0.66$) compared to other subjects ($0.18 \leq r_{average} \leq 0.20$) for the offer contrast map, $T(125) \geq 17.6$, $p < .001$. The difference is visible in the prominent diagonal in the color-coded matrices (see left panel Fig. 6). In addition, the blue empirical cumulative density function is shifted to the right, depicting higher within-subject similarities compared to between-subject similarities in

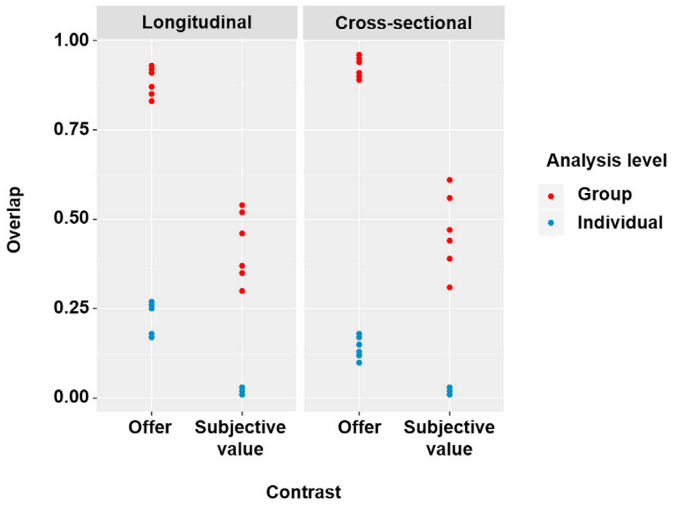

Fig. 5. Overlap is lower for the subjective-value contrast compared to the offer contrast (columns) and lower at the individual level compared to the group level (colors), but similar for longitudinal and cross-sectional analyses (panels). Each dot depicts an overlap coefficient separated for longitudinal and cross-sectional analysis, group and individual, and offer vs. subjective-value contrasts.

red. Although the difference between similarity within subjects ($0.07 \leq r_{average} \leq 0.08$) compared to other subjects ($r_{average} = 0.01$) was also significant for the subjective-value contrast, $T(125) \geq 2.1$, $p < .037$, the difference is not as prominent as in the offer contrast (see right panel Fig. 6). Moreover, the within-subject similarity was significantly higher for the offer contrast compared to the subjective-value contrast, $T(125) \geq 10.0$, $p < .001$. Analogous results were obtained for cross-sectional reliability using the split-half method (see Table S.2 and S.3 for complete T-test results).

In the next step, we examined whether we could re-identify individuals based on maximum similarity (Finn et al., 2015). Again, there was a clear difference between the offer contrast and the subjective-value contrast. For the offer contrast, 84 up to 126 (all) participants (67%–100% ranging across comparisons) could be re-identified based on their maximum similarity to a second scan, whereas it was only the case for 8 to 18 participants (6%–14%) for the subjective-value contrast (Fig. 6). Note that according to a binomial distribution, ≥3 correct classifications would be considered as better than chance ($p < .0185$) indicating that both contrasts work significantly better than chance in re-identifying individuals.

To further disentangle sources of variability, variance analyses with the factors subject level (within vs. between) and contrast (subjective value vs. offer) revealed a significant interaction between subject level and contrast for all comparisons, $Fs \geq 147.9$, $ps < .001$ (Fig. 7, Table S.1 for complete ANOVA results). In general, similarities were lower for the subjective-value contrast than the offer contrast, $Fs \geq 659.7$, $ps < .001$. Thus, the difference between within- and between-subject similarities is higher for the offer contrast indicating that the signal elicited during the offer contrast contains more unique individual information compared to the subjective-value contrast.

To provide another reference for basic reliability of task-evoked brain response, we assessed similarities for the motor contrasts. This analysis yielded highly comparable results to the offer contrast: substantially higher within-subject compared to between-subject similarity (Figure S.10). Lastly, we ran simulations with known signal (i.e.,

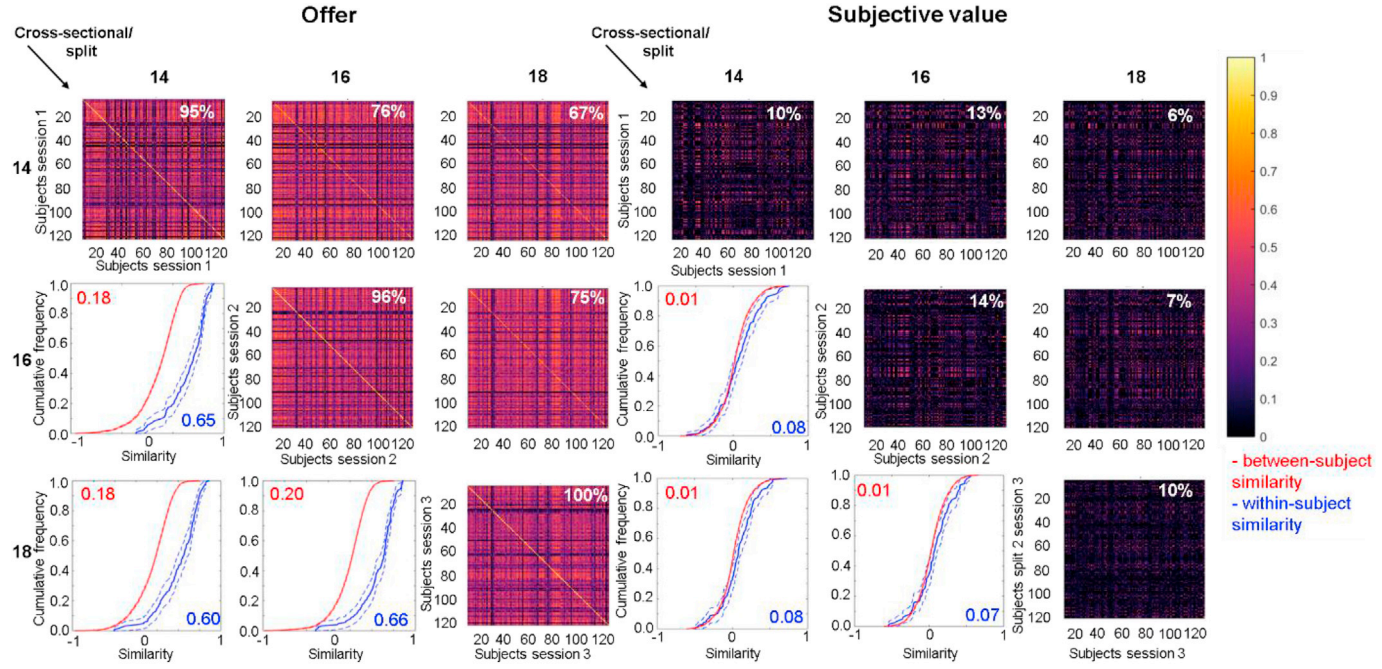

Fig. 6. Similarity results for all longitudinal and cross-sectional comparisons for the offer contrast (left panel) and subjective-value contrast (right panel). In each panel, the color-coded maps in the upper diagonal represent the longitudinal comparisons, while the cross-sectional comparisons are visible in the diagonal. In the color-coded maps, each row depicts the correlation of the activation map of one person for one session or half of a session (split) with other persons in the sample (in columns). For reliable activation "fingerprints", we would expect the highest similarity with activation maps coming from the same person (within-subject similarity). This would be visible in a prominent diagonal as is the case for the offer contrast (left panel). The numbers printed on the similarity matrices indicate the percentage of correctly re-identified persons. Participants were correctly re-identified if they were most similar to a different session or split coming from the same person. Re-identification was considerably higher for the offer contrast. This difference was also evident in the empirical cumulative density functions (ecdf; lower diagonal of each panel; for details see methods): The higher the similarity is on average (depicted in upper left and lower right corner of ecdf), the more the lines will be shifted to the right (red: between subject, blue: within subject). Note that there is little person-specific similarity contained in the subjective-value contrast which is indicated by the overlap of the colored lines. The dotted lines represent the 95% confidence interval. The data used to draw this figure is provided in the Supporting Information.

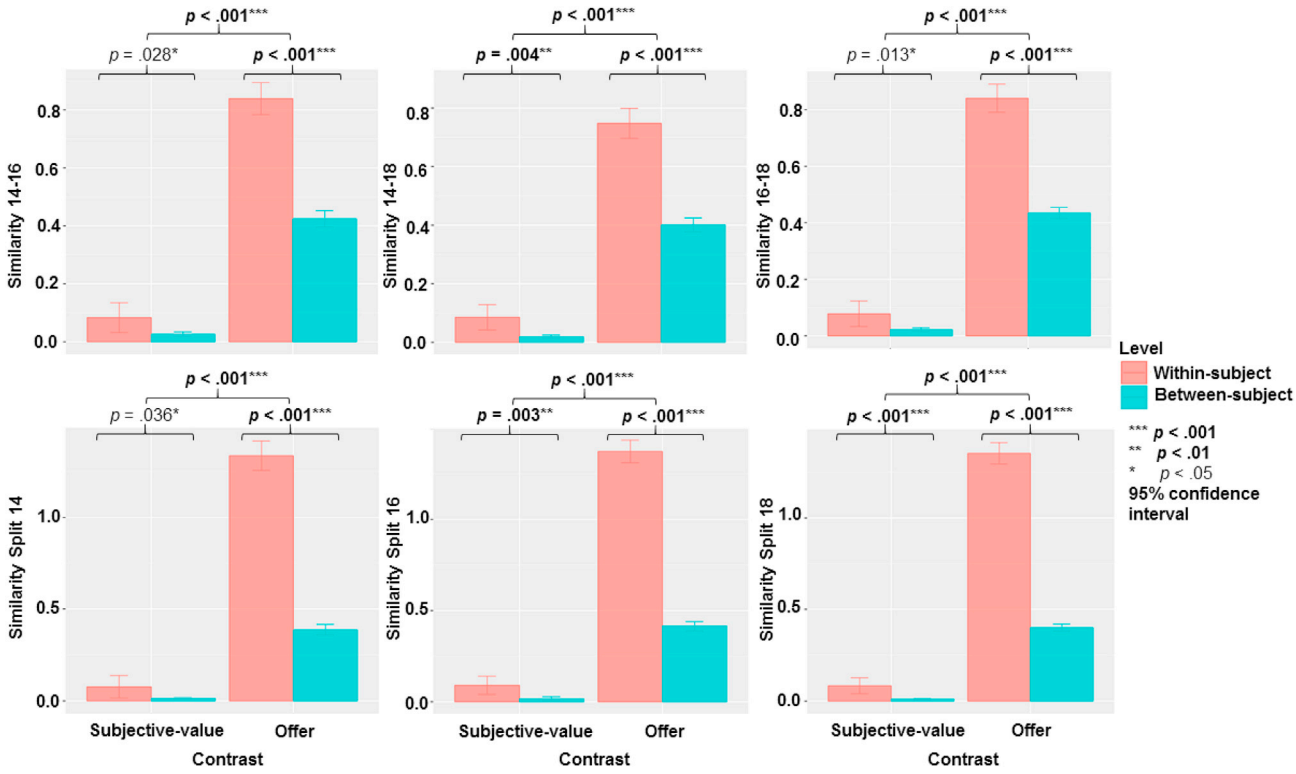

Fig. 7. Summary of z-transformed similarities for each comparison. Stars indicate significance of paired T-tests for within- and between-subject similarities (offer: $T \geq 17.6$, $p \leq .001$; subjective value: $T \geq 2.1$, $p \leq .036$) and for the difference between within- and between-subject similarities. The offer contrast was characterized by a greater difference in within-vs. between-subject similarity compared to the subjective-value contrast ($T \geq 10.0$, $p \leq .001$; see Supporting Information).

individual activation maps at age 14), which were corrupted by noise mimicking changes over time primarily due to measurement error. Even for the highest level of noise added to the individual contrast maps (i.e., double the initial variance), within-subject reliability was preserved to a moderate extent suggesting that the absence of within-subject reliability is more than a simple result of noisy test-retest data (Fig. 8).

To conclude, the task elicits reliable brain activation patterns that can be used to re-identify participants with high accuracy. Yet, the subjective-value contrast fails to achieve the required minimum level of reliability. This suggests that the low reliability is not due to the collected data, the

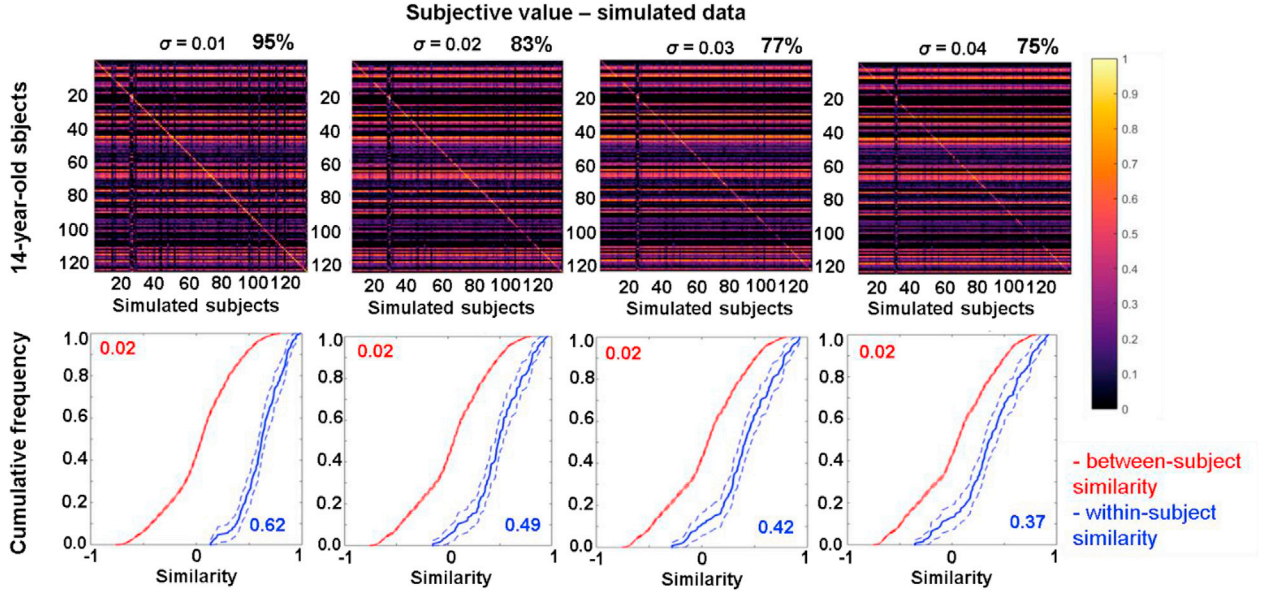

Fig. 8. Similarity maps and empirical cumulative distribution functions (red: between-subject; blue: within-subject) for simulated changes in the subjective-value contrast based on brain activation patterns of 14-year-olds and with increasing noise from $\sigma = 0.01$ (half the initial variance) on the left to $\sigma = 0.04$ (double the initial variance) on the right. Whereas increasing levels of noise in the individual estimates led to less within-subject similarity (diagonal in the matrices and blue lines in cumulative frequency plots), excessive measurement noise alone cannot explain the low reliability that we empirically observed in the subjective-value contrast. The numbers printed on the upper right of the similarity matrices indicate the percentage of correctly re-identified subjects based on maximum similarity (cf. Fig. 6).

task, or high measurement noise *per se*, but rather attributable to the specific parametric contrast intended to track subjective value.

3.4. Local reliability

Next, we sought to identify regional differences in the reliability of task-evoked brain responses by calculating ICCs and correlations for each voxel. Across the brain, ICCs and correlations were much lower for the subjective-value contrast than the offer contrast. Results were also highly similar for the different longitudinal and cross-sectional correlation coefficients (Fig. 9). Regional differences in reliability were visible for the offer contrast where we observed higher correlations in visual and parietal regions and lower correlations in orbitofrontal regions. Notably, there was no such apparent pattern of regional differences in the reliability of the subject-value contrast. Even in the commonly identified value-tracking regions, we found higher average reliability for the offer contrast compared to the subjective-value contrast (for unthresholded maps and corresponding between- and within-subject variances, see https://neurovault.org/collections/KEASERVU/).

To summarize voxel-wise data according to neuroanatomy, we created a matrix with an average ICC and correlation for each region included in the CONN atlas (Whitfield-Gabrieli and Nieto-Castanon, 2012). For visualization purposes, we grouped the ROIs in functional networks according to Yeo et al. (2011). Again, we observed that the reliability of the offer contrast was higher compared to the subjective-value contrast (Fig. 10). Moreover, network-based differences in reliability occurred for the offer contrast (highest in the visual network), but there was no conclusive indication of network-based differences in reliability for the subjective-value contrast. The correlation of split-half reliabilities of both contrasts did reveal a significant, but weak positive association ($r = 0.14$, $p < .001$; Fig. 10). Notably, reliability only surpassed the moderate criterion for both contrasts in the nucleus accumbens in 18-year-old participants.

3.4.1. Statistical dependence of reliability and signal amplitude and variance

So far, we have shown that the offer contrast outperforms the subjective-value contrast in global and local reliability. This low reliability of the subjective-value contrast may seem surprising given the congruent group main effects of the contrast. Hence, we analyzed the correspondence of reliability with contrast amplitude and its variance at the group level as well as the ratio between the two that is used to calculate the t-statistic. Our rationale was to test if the scaling between these key metrics is different among contrasts or if the observed differences in reliability are merely due to changes in average contrast amplitude. To this end, within each gray-matter voxel, we correlated the split-half reliability, Pearson's *r*, of each session with the average amplitude and variance of the contrast across participants.

In line with the expected association, we found that higher average betas were associated with greater split-half reliability for the offer contrast (Spearman's rho = .55, $p < .001$), but only to a negligible extent for the subjective-value contrast (Spearman's rho = 0.06, $p < .001$). Further in line with our hypothesis that greater inter-individual variability is important for reliability at the individual level, we observed a positive association between the two for the offer contrast (rho = 0.57, $p < .001$) and, to a weaker extent, for the subjective-value contrast as well (rho = 0.18, $p < .001$). Perhaps surprising at first, this led to an attenuated rank-order correlation for t-values with reliability (offer: rho = 0.45, $p < .001$; subjective value: rho = 0.00, $p = .43$; Fig. 11) compared to average amplitude or inter-individual variability alone. Moreover, we conducted an additional exploratory analysis where we removed the shared variance with contrast amplitude from the inter-individual variability of the contrast (orthogonalization) using linear regression. This analysis showed higher reliability in voxels with lower or higher inter-individual variability in brain response than expected based on the contrast amplitude alone (Fig. 12). Taken together, these results indicate that a substantial degree of the reliability in contrasts that show good psychometric characteristics can be accounted for by signaling characteristics within contrasts.

3.4.2. fMRI reliability and behavioral consistency

Since task-evoked fMRI signal should reflect behavior, the consistency of the behavior might influence the reliability of the corresponding brain signal. Therefore, we tested whether behavioral consistency, operationalized as β of the softmax function or BIC, is related to brain

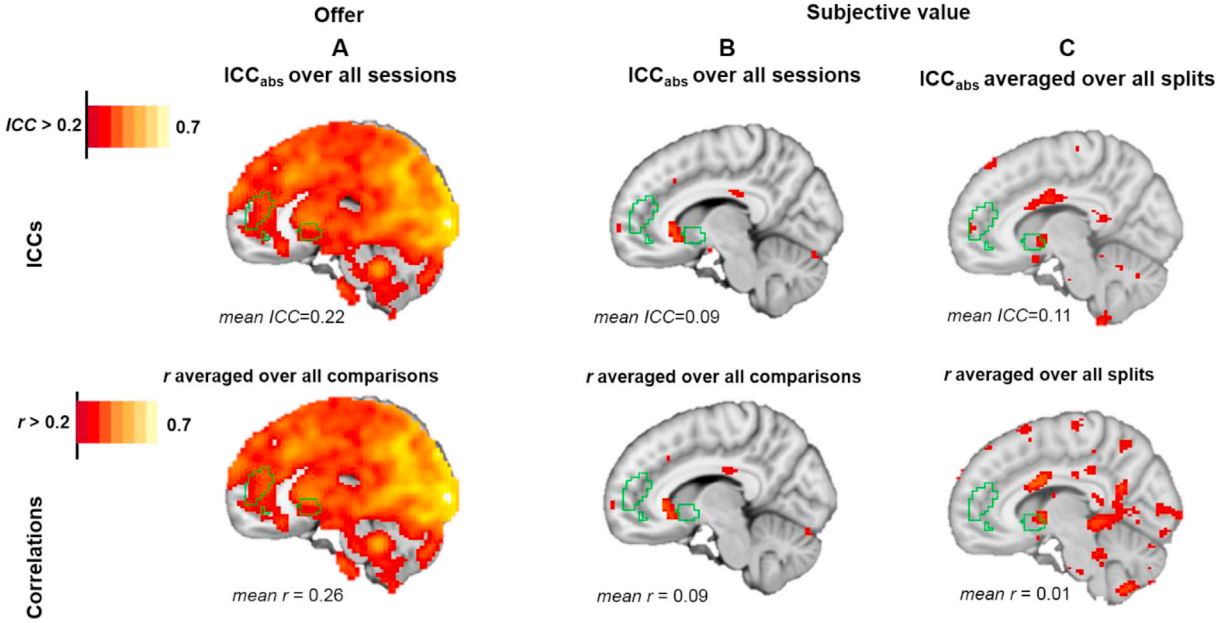

Fig. 9. Absolute intra-class correlation coefficient *(ICC_{abs})* and Pearson's correlation *(r)* over all sessions (longitudinal) depicted for a selected sagittal slice for (A) the offer contrast, (B) the subjective-value contrast and (C) for the averaged split (cross-sectional) data for the subjective-value contrast. Average values indicate reliability within the independently identified ROI encompassing ventral striatum and vmPFC (green outline).

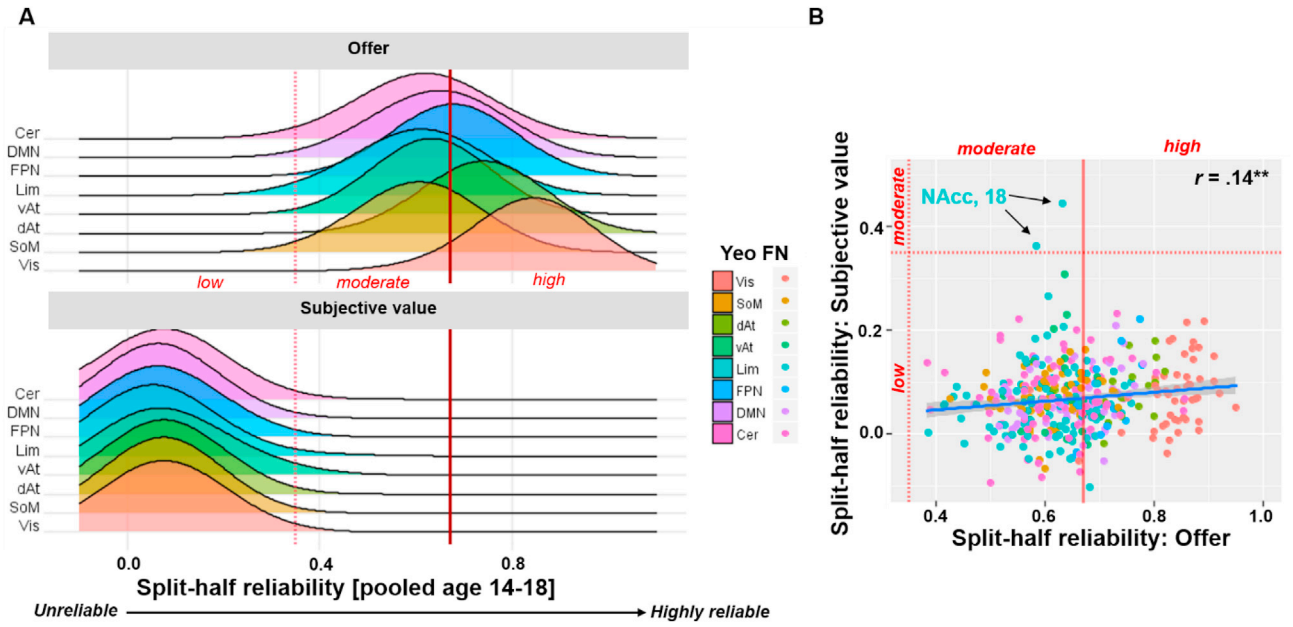

Fig. 10. (A) Density plots of the split-half reliabilities, *r*, for offer and subjective-value contrasts over all assessments. (B) Correlation between split-half reliabilities, *r*, for offer and subjective-value contrasts in each region listed in the atlas. Reliability for both contrasts only exceeded the moderate criterion within the nucleus accumbens (NAcc) of 18-year-old participants. (A + B) colored according to Yeo et al.'s (2011) functional networks (FN): visual (Vis), somatomotor (SoM), dorsal attention (dAt), ventral attention (vAt), limbic (lim; including subcortical limbic regions), frontoparietal (FPN), default mode (DMN), cerebellum (Cer; added as network). Each dot represents one region in the atlas.

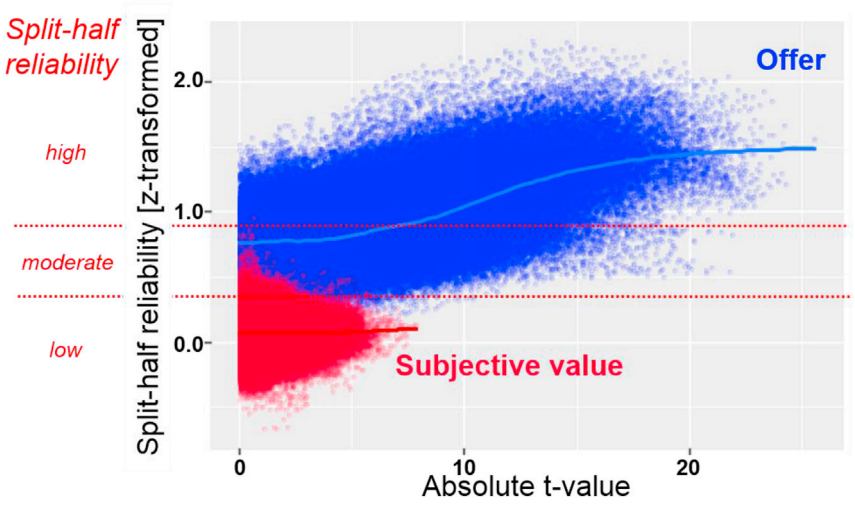

Fig. 11. Association of absolute t-values and split-half reliability for offer (blue) and subjective value (red). Note that for t-values of comparable magnitude, voxel-wise reliability is lower in the subjective-value contrast compared to the offer contrast. Colored smoothing lines were calculated via the ggplot2 'generalized additive models with integrated smoothness estimation' (gam) option. Each dot depicts a voxel.

response reliability or the amplitude of a contrast *per se*. Therefore, we conducted a hierarchical linear model analysis to test whether behavioral choice consistency can predict (1) fMRI reliability measured via within- and between-subject similarity and (2) the contrast estimates for the offer and subjective-value contrast. Hierarchical linear modeling showed that both, the consistency parameter β and the model fit BIC, were associated with within-subject similarity in both contrasts: higher choice consistency was associated with higher within-subject similarity. Notably, there was a positive association of choice consistency with between-subject similarity in the offer contrast, but not in the subjective-value contrast (Fig. 13). However, there were no significant associations between consistency and contrast amplitude (Figure S.11).

4. Discussion

Reliability is a key aspect of the diagnostic quality of any measurement such as a biomarker derived from brain activation to monetary offers. Yet, despite the recent surge of interest in biomarkers, little is known about the reliability of many paradigms commonly used in fMRI research. Here, we provided a comprehensive approach to investigate cross-sectional (within-session) and longitudinal (between-session) reliability of fMRI data. Using longitudinal data of an intertemporal choice task, we described an extensive analysis of reliability spanning from the group to the individual level and from the global to the local level. The key results were consistent across all applied measures: First, the

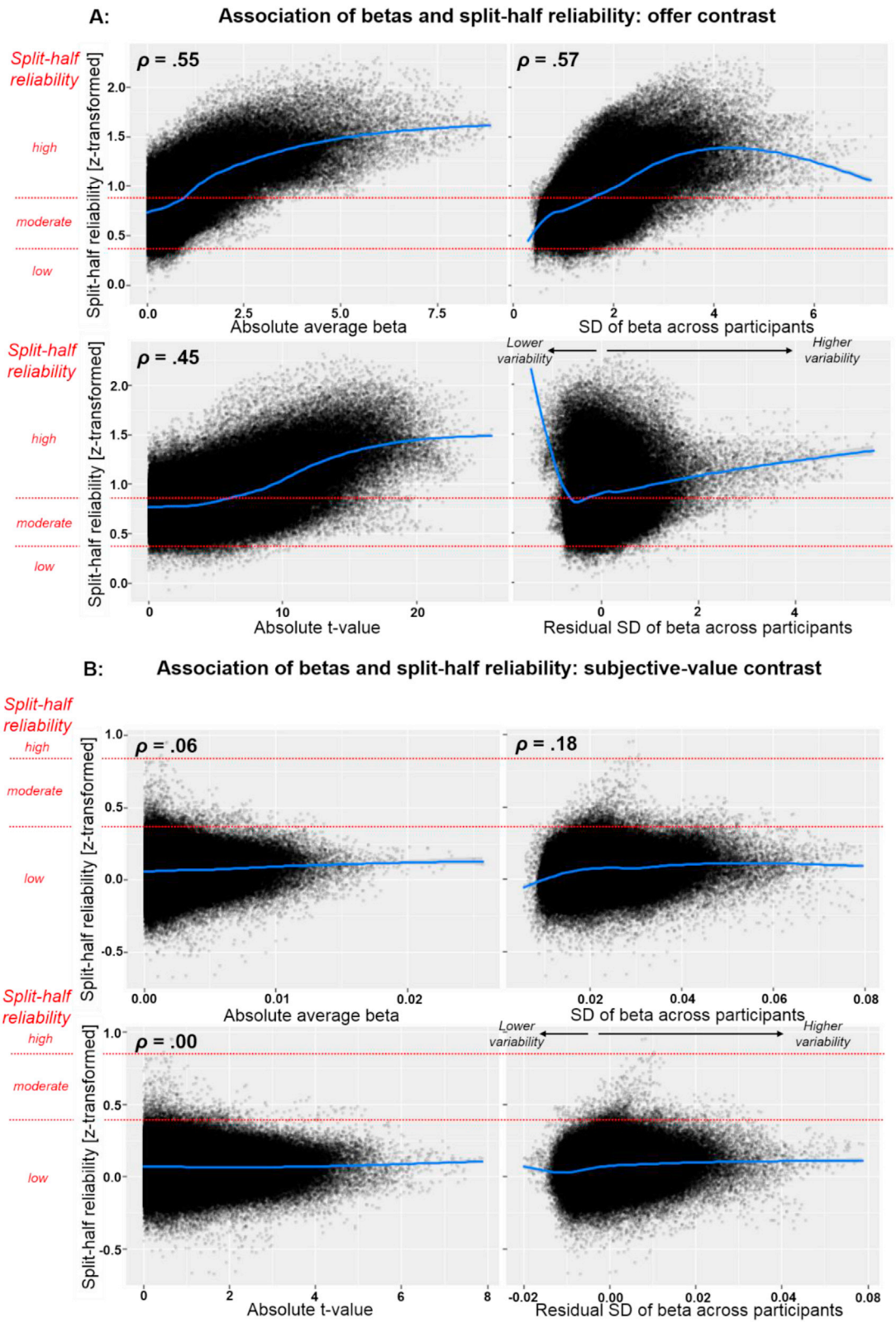

Fig. 12. Moderately strong associations between contrast amplitude (beta and t-value), inter-individual variability (SD and residual SD) and split-half reliability for the offer contrast (A), but only weak associations for the subjective-value contrast (B). Colored smoothing lines were calculated via the ggplot2 'generalized additive models with integrated smoothness estimation' (gam) option. Note that residual SD depicts lower or higher variability according to a linear regression analysis using contrast amplitude as predictor. Each dot depicts a voxel.

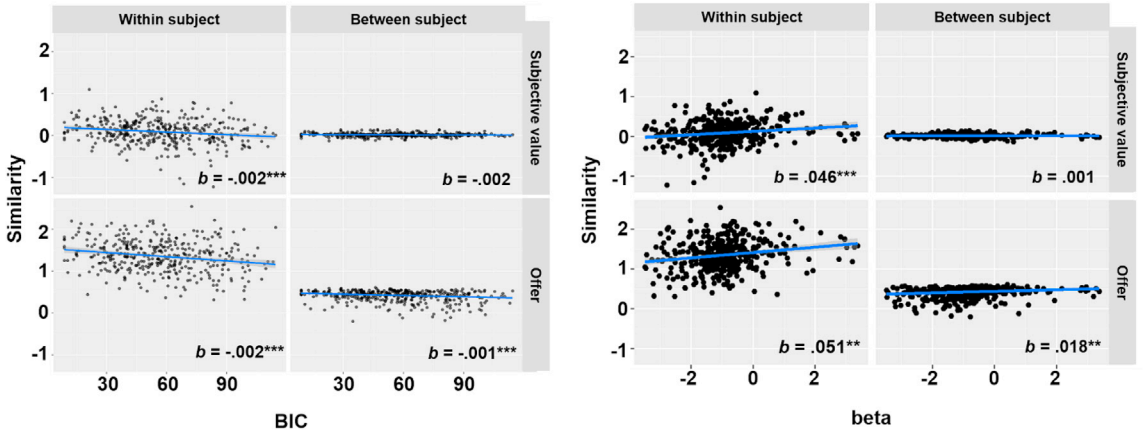

Fig. 13. Higher choice consistency is associated with higher within-subject similarity. Association between two consistency measures (on the left: the Bayesian Information Criterion (BIC) as goodness-of-fit of the hyperbolic discounting function and on the right: β from the hyperbolic discounting equation (Ripke et al., 2012) and similarities, assessed for both contrasts as well as within and between subjects. Note that higher BIC scores represent lower model fit and therefore lower consistency. Blue lines show robust fit lines; beta (*b,* lower right corner in each graph) values indicate the slopes (***: $p < .001$, **: $p < .01$). Each dot depicts one session per participant. Mixed-effects models account for the nested structure of the data and calculate *p*-values at the level of participants.

individual reliability of the brain response to the offer onset is substantially higher than the reliability of the parametric value-tracking signal. Second, group-level reliability is substantially higher than individual reliability. Third, cross-sectional and longitudinal reliability of the subjective-value contrast were similarly low and failed to achieve a minimum level of reliability that is required for valid use as a biomarker. Fourth, we provide preliminary evidence that reliability is positively dependent on inter-individual variability in addition to contrast amplitude. Collectively, our results emphasize the necessity to optimize between-subject variance for reliable classification and prediction.

4.1. Good reliability at the group level versus unreliability at the individual level

In line with previous studies on the stability of group effects (Bennett and Miller, 2010; Freyer et al., 2009; Nord et al., 2017; Plichta et al., 2012), we observed congruent brain activation for both onset and subjective value of the delayed offer over time. However, individual trajectories were not reliable over time as well as inconsistent across split halves of each session for the subjective-value contrast. Since our study is not the first revealing such notable difference between individual- and group-level reliability (van den Bulk et al., 2013; Vetter et al., 2015), these results call for caution in using group-level results to draw conclusions about individual aspects of brain function (Fisher et al., 2018). Illustratively, increased group-level brain activity in a specific region during the choice of one option does not necessarily imply that these choice "signatures" would suffice to predict choice preferences at the individual level, which depends on individual reliability (Fisher et al., 2018). Thus, our results might explain the limited success in connecting neural and behavioral results for many commonly employed tasks in cognitive neuroscience (Müller et al., 2015; Nebe et al., 2018; Whelan et al., 2014), perhaps due to an emphasis on paradigms producing robust main effects (e.g. Hedge et al., 2018).

Whereas differences between the individual versus the group level were striking, multiple sources are likely to contribute to the observed pattern. For example, greater measurement noise contained in the individual versus the averaged group-level brain response patterns could be a factor. However, in our simulation, we doubled the inter-individual variability of the first-wave estimates in subjective-value tracking and were still able to recover a considerably more reliable rank order compared to the actual reliability of the parametric contrast. Moreover, good longitudinal reliability and whole-brain similarity, as were evident in the offer or motor contrasts, speak against a simple measurement noise account. Hence, our reliability estimates obtained from simple event contrasts do not support a holistic concern that fMRI data in general is too noisy to be used as a biomarker for individualized prediction in clinical research. Nevertheless, it emphasizes that not all contrasts are equally useful for prediction and classification at an individual level, at least according to their basic psychometric characteristics. Similarly, our results demonstrate that not only the amplitude of the brain signal but also the inter-individual variance in the contrast is positively associated with its reliability. Hence, inter-individual variance is needed to capture brain activation as an individual characteristic and to differentiate it from a commonly shared neural evaluation process (Hedge et al., 2018).

4.2. Use of reliable brain activation patterns as a potential tool for clinical research

In the search for potential biomarkers of mental disorders, reliability at an individual level is a major limitation in gauging its potential predictive value (Nord et al., 2017). Hence, we should expect a notably higher within-subject compared to between-subject similarity for any promising candidate such as the value-tracking contrast (Ripke et al., 2014). In our case, the within-subject similarity was indeed higher than the between-subject similarity. However, the absolute values and the difference between individual and group levels was much lower for the subjective-value contrast and largely failed to exceed the recommended minimum threshold for moderate reliability. Whereas developmental changes could partly explain low longitudinal reliability occurring between sessions, such changes cannot explain the low consistency across split halves of the paradigm within a given session. Also, the high intra-individual similarity and correspondingly high accuracy in re-identifying participants using brain activation patterns derived from the offer contrast suggests that there is a distinct individual component contained in the processing of monetary offers. Intriguingly, the congruent group activation in the subjective-value contrast might indicate that the underlying value-tracking signal commonly seen in the ventral striatum and vmPFC could be immutably shared among participants. This would explain why an unreliable individual response occurs within a well-replicated network consistently shown at the group level (McClure et al., 2004; Ripke et al., 2012), which is also conclusively supported by other neuroscientific methods including decades of animal research (e.g., Floresco, 2013; Floresco et al., 2008; Hamid et al., 2016; Saddoris et al., 2015). Taken together, these results suggest that the parametric contrast for subjective value is not a suitable candidate as a biomarker for individualized prediction because it fails basic diagnostic criteria, whereas the offer contrast achieved a sufficiently high reliability across individuals and sessions. Results from our group support this

interpretation, since we found associations with intelligence (Ripke et al., 2014) and smoking status (Kobiella et al., 2014) for the offer/decision phase in intertemporal choice tasks. However, we failed to find differences between smokers and healthy controls and observed no effects of smoking cessation in the subjective-value contrast with this task (Grosskopf et al., in prep). Furthermore, to the best of our knowledge, there are also no conclusively replicated associations between subjective-value signals derived from intertemporal choice and clinical characteristics to date although the temporal discounting rate k has been associated with many diverse health-related outcomes (Story et al., 2014).

More generally, a weak correspondence between the estimated brain response in two independent halves, runs, or sessions indicates a failure in reliably differentiating between individuals. In turn, this makes it improbable to detect associations with other more distant outcomes (Hedge et al., 2018). During the last decade, several large-scale studies started to investigate the predictive validity of fMRI data for select aspects of human behavior such as the IMAGEN project (Schumann et al., 2010), the UK biobank (www.ukbiobank.ac.uk; Sudlow et al., 2015) or the Human Connectome Project (Van Essen et al., 2012). Whereas the predictive validity of many classic paradigms used in cognitive neuroscience may turn out to be limited (Hedge et al., 2018), these large studies offer the possibility to investigate such paradigms systematically regarding psychometric characteristics. Ultimately, this may help in identifying paradigms, contrasts, and conditions yielding a sufficiently high reliability for prediction and classification of individuals. To this end, our toolbox fmreli may facilitate the development of a standardized and comprehensive approach in establishing the reliability of biomarkers derived from fMRI data. Specifically, our results call for greater attention to identify sufficiently reliable estimates of subjective-value tracking across the set of common tasks and brain response characteristics to improve the potential of value-tracking correlates as future biomarkers for mental disorders.

4.3. Limitations

Our study and the reliability analyses provided here have several limitations. First, within-run reliability is not an ideal measure as it might be biased to some degree due to, for example, higher statistical dependence. However, not every study has data from at least two different sessions or runs available and the design simulation function in the toolbox can be used to safeguard against unfounded conclusions. Preferably, independent runs should be collected to initially establish the reliability of a task, but we demonstrate that within-run estimates can provide useful insights when multiple runs are not available or were collected years apart. Second, the current study was limited to one representative paradigm. Thus, the results may not generalize to other tasks. Still, the intertemporal choice paradigm provides a striking case as it is often used in fMRI research and we had sufficient data across three waves of data collection (i.e., >1 h of task-based fMRI data per participant). The high consistency of the obtained reliability estimates across waves further corroborates the evidence provided by this key example for other relevant scenarios. Thus, we feel that our results warrant calling for more caution in future research targeting individual aspects of brain function, particularly in longitudinal studies. Further investigations may contribute to specific aspects of the design that determine individual reliability of more nuanced facets of value tracking. Third, in our case, we used several measures to facilitate comparison and combination of different measures. So far, we primarily collected the most commonly employed reliability indices in our new toolbox to make them more readily accessible for fMRI research. However, new methods for the estimation of fMRI reliability have been proposed that may provide additional insights and should be implemented in the future to make better use of the unique characteristics of fMRI data (Maitra et al., 2002; Shou et al., 2013; Zandbelt et al., 2008). To this end, we welcome the future addition of novel measures to the toolbox to accelerate the dissemination among the fMRI community. Fourth, other statistical methods incorporating hierarchical priors on parameter distributions that improve the recovery of brain response estimates could be employed in the future to improve the accuracy of individual estimates (e.g. Kroemer et al., 2014, 2016; Mejia et al., 2018).

5. Conclusions

Using an intertemporal choice task, we have shown that there is a substantial difference between group-level and individual-level reliability in brain response to monetary offers and that the extent of that difference varies strongly between contrasts within the task. Simple contrasts reflecting activation elicited by the presentation of monetary offers or motor responses showed good reliability and allowed us to re-identify individuals with high accuracy across multiple waves. Critically, the subjective-value contrast, which is commonly used to assess individual differences in value tracking showed insufficiently low reliability across multiple indicators and levels of analysis. To conclude, our results suggest that promising biomarkers should be extensively evaluated with respect to intra-individual stability over time before they can be routinely applied for prediction or classification to avoid the reliability fallacy arising from congruent group activation maps. Importantly, we provide all the functions that we have employed in the MATLAB-based toolbox fmreli to facilitate the use in future analyses of fMRI reliability across many more applications in cognitive neuroscience.

Code availability

The toolbox fmreli is available via GitHub (https://github.com/nkroemer/reliability) and further documentation is available on https://neuromadlab.com/resources/fmreli. In addition, the ICC maps (absolute agreement) and the respective variance components are provided on NeuroVault https://neurovault.org/collections/KEASERVU/.

Acknowledgements

This study was supported by the Deutsche Forschungsgemeinschaft (DFG Grant # SFB 940/1 & SFB 940/2 and the German Ministry of Education and Research (BMBF Grant # 01EV0711 & # 01EE1406B). JHF received a PhD-scholarship from the SFB 940 „Volition and Cognitive Control: mechanisms, modulators and dysfunctions". VT and NBK were supported by the University of Tübingen's Faculty of Medicine fortüne program, grant #2453-0-0.

We thank Stephan Ripke for his previous work on the presented intertemporal choice task. We thank Marie Stolze and Caroline Burrasch, who contributed to the implementation and documentation of the toolbox.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.03.053.

References

Aron, A.R., Gluck, M.A., Poldrack, R.A., 2006. Long-term test–retest reliability of functional MRI in a classification learning task. Neuroimage 29 (3), 1000–1006.

Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Ann. N. Y. Acad. Sci. 1191 (1), 133–155. https://doi.org/10.1111/j.1749-6632.2010.05446.x.

Bickel, W.K., Pitcock, J.A., Yi, R., Angtuaco, E.J.C., 2009. Congruence of BOLD response across intertemporal choice conditions: fictive and real money gains and losses. J. Neurosci. 29 (27), 8839–8846. https://doi.org/10.1523/JNEUROSCI.5319-08.2009.

Caceres, A., Hall, D.L., Zelaya, F.O., Williams, S.C.R., Mehta, M.A., 2009. Measuring fMRI reliability with the intra-class correlation coefficient. Neuroimage 45 (3), 758–768. https://doi.org/10.1016/j.neuroimage.2008.12.035.

Cicchetti, D.V., Sparrow, S.A., 1981. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am. J. Ment. Defic. 86 (2), 127–137.

Cicchetti, D.V., 2001. The precision of reliability and validity estimates re-visited: distinguishing between clinical and statistical significance of sample size requirements. J. Clin. Exp. Neuropsychol. 23 (5), 695.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., et al., 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31 (3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021.

Dubois, J., Adolphs, R., 2016. Building a science of individual differences from fMRI. Trends Cognit. Sci. 20 (6), 425–443. https://doi.org/10.1016/j.tics.2016.03.014.

Faul, F., Erdfelder, E., Lang, A.-G., Buchner, A., 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods 39 (2), 175–191.

Finn, E.S., Scheinost, D., Finn, D.M., Shen, X., Papademetris, X., Constable, R.T., 2017. Can brain state be manipulated to emphasize individual differences in functional connectivity? Neuroimage 160, 140–151. https://doi.org/10.1016/j.neuroimage.2017.03.064.

Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., et al., 2015. Functional connectome fingerprinting: identifying individuals based on patterns of brain connectivity. Nat. Neurosci. 18 (11), 1664–1671. https://doi.org/10.1038/nn.4135.

Fisher, A.J., Medaglia, J.D., Jeronimus, B.F., 2018. Lack of group-to-individual generalizability is a threat to human subjects research. Proc. Natl. Acad. Sci. Unit. States Am. 115 (27), E6106–E6115. https://doi.org/10.1073/pnas.1711978115.

Fleiss, J.L., 1986. The Design and Analysis of Clinical Experiments. Wiley, NY.

Fliessbach, K., Rohe, T., Linder, N.S., Trautner, P., Elger, C.E., Weber, B., 2010. Retest reliability of reward-related BOLD signals. Neuroimage 50 (3), 1168–1176. https://doi.org/10.1016/j.neuroimage.2010.01.036.

Floresco, S.B., 2013. Prefrontal dopamine and behavioral flexibility: shifting from an "inverted-U" toward a family of functions. Front. Neurosci. 7. https://doi.org/10.3389/fnins.2013.00062.

Floresco, S.B., Maric, T.L., Ghods-Sharifi, S., 2008. Dopaminergic and glutamatergic regulation of effort-and delay-based decision making. Neuropsychopharmacology 33 (8), 1966.

Frazier, J.A., Chiu, S., Breeze, J.L., Makris, N., Lange, N., Kennedy, D.N., et al., 2005. Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. Am. J. Psychiatry 162 (7), 1256–1265. https://doi.org/10.1176/appi.ajp.162.7.1256.

Freyer, T., Valerius, G., Kuelz, A.-K., Speck, O., Glauche, V., Hull, M., Voderholzer, U., 2009. Test–retest reliability of event-related functional MRI in a probabilistic reversal learning task. Psychiatr. Res. Neuroimaging 174 (1), 40–46. https://doi.org/10.1016/j.pscychresns.2009.03.003.

Garrett, D.D., Samanez-Larkin, G.R., MacDonald, S.W.S., Lindenberger, U., McIntosh, A.R., Grady, C.L., 2013. Moment-to-moment brain signal variability: a next frontier in human brain mapping? Neurosci. Biobehav. Rev. 37 (4), 610–624. https://doi.org/10.1016/j.neubiorev.2013.02.015.

Gee, D.G., McEwen, S.C., Forsyth, J.K., Haut, K.M., Bearden, C.E., Addington, J., et al., 2015. Reliability of an fMRI paradigm for emotional processing in a multisite longitudinal study. Hum. Brain Mapp. 36 (7), 2558–2579. https://doi.org/10.1002/hbm.22791.

Goldstein, J.M., Seidman, L.J., Makris, N., Ahern, T., O'Brien, L.M., Caviness, V.S., et al., 2007. Hypothalamic abnormalities in schizophrenia: sex effects and genetic vulnerability. Biol. Psychiatry 61 (8), 935–945. https://doi.org/10.1016/j.biopsych.2006.06.027.

Gorgolewski, K.J., Storkey, A.J., Bastin, M.E., Whittle, I., Pernet, C., 2013. Single subject fMRI test–retest reliability metrics and confounding factors. Neuroimage 69, 231–243. https://doi.org/10.1016/j.neuroimage.2012.10.085.

Groppe, D.M., Makeig, S., Kutas, M., 2009. Identifying reliable independent components via split-half comparisons. Neuroimage 45 (4), 1199–1211. https://doi.org/10.1016/j.neuroimage.2008.12.038.

Großkopf, C.M., Kroemer, N.B., Pooseh, S., Böhme, F., Smolka, M.N., Temporal discounting and choice consistency in smoking cessation: links from a longitudinal fMRI study, submitted manuscript.

Hamid, A.A., Pettibone, J.R., Mabrouk, O.S., Hetrick, V.L., Schmidt, R., Vander Weele, C.M., et al., 2016. Mesolimbic dopamine signals the value of work. Nat. Neurosci. 19 (1), 117–126. https://doi.org/10.1038/nn.4173.

Hare, T.A., Camerer, C.F., Rangel, A., 2009. Self-control in decision-making involves modulation of the vmPFC valuation system. Science 324 (5927), 646–648. https://doi.org/10.1126/science.1168450.

Havel, P., Braun, B., Rau, S., Tonn, J.-C., Fesl, G., Brückmann, H., Ilmberger, J., 2006. Reproducibility of activation in four motor paradigms. J. Neurol. 253 (4), 471–476. https://doi.org/10.1007/s00415-005-0028-4.

Hedge, C., Powell, G., Sumner, P., 2018. The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. Behav. Res. Methods 50 (3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1.

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

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., et al., 2010. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am. J. Psychiatry 167 (7), 748–751. https://doi.org/10.1176/appi.ajp.2010.09091379.

Jaccard, P., 1901. Ètude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin del la Sociètè Vaudoise des Sciences Naturelles 37, 547579.

Jurk, S., Mennigen, E., Goschke, T., Smolka, M.N., 2018. Low-level alcohol consumption during adolescence and its impact on cognitive control development. Addict. Biol. 23 (1), 313–326. https://doi.org/10.1111/adb.12467.

Kable, J.W., Glimcher, P.W., 2007. The neural correlates of subjective value during intertemporal choice. Nat. Neurosci. 10 (12), 1625–1633. https://doi.org/10.1038/nn2007.

Kobiella, A., Ripke, S., Kroemer, N.B., Vollmert, C., Vollstädt-Klein, S., Ulshöfer, D.E., Smolka, M.N., 2014. Acute and chronic nicotine effects on behaviour and brain activation during intertemporal decision making. Addict. Biol. 19 (5), 918–930. https://doi.org/10.1111/adb.12057.

Koffarnus, M.N., Deshpande, H.U., Lisinski, J.M., Eklund, A., Bickel, W.K., LaConte, S.M., 2017. An adaptive, individualized fMRI delay discounting procedure to increase flexibility and optimize scanner time. Neuroimage 161 (Suppl. C), 56–66. https://doi.org/10.1016/j.neuroimage.2017.08.024.

Koolschijn, P.C.M.P., Schel, M.A., Rooij, M. de, Rombouts, S.A.R.B., Crone, E.A., 2011. A three-year longitudinal functional magnetic resonance imaging study of performance monitoring and test-retest reliability from childhood to early adulthood. J. Neurosci. 31 (11), 4204–4212. https://doi.org/10.1523/JNEUROSCI.6415-10.2011.

Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational similarity analysis – connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2. https://doi.org/10.3389/neuro.06.004.2008.

Kroemer, N.B., Guevara, A., Ciocanea Teodorescu, I., Wuttig, F., Kobiella, A., Smolka, M.N., 2014. Balancing reward and work: anticipatory brain activation in NAcc and VTA predict effort differentially. Neuroimage 102 (2), 510–519. https://doi.org/10.1016/j.neuroimage.2014.07.060.

Kroemer, N.B., Sun, X., Veldhuizen, M.G., Babbs, A.E., de Araujo, I.E., Small, D.M., 2016. Weighing the evidence: variance in brain responses to milkshake receipt is predictive of eating behavior. Neuroimage 128, 273–283. https://doi.org/10.1016/j.neuroimage.2015.12.031.

Loubinoux, I., Carel, C., Alary, F., Boulanouar, K., Viallard, G., Manelfe, C., et al., 2001. Within-session and between-session reproducibility of cerebral sensorimotor activation: a test–retest effect evidenced with functional magnetic resonance imaging. J. Cereb. Blood Flow Metab. 21 (5), 592–607. https://doi.org/10.1097/00004647-200105000-00014.

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. https://doi.org/10.1016/j.neuroimage.2009.11.070.

Maitra, R., Roys, S.R., Gullapalli, R.P., 2002. Test-retest reliability estimation of functional MRI data. Magn. Reson. Med. 48 (1), 62–70. https://doi.org/10.1002/mrm.10191.

Makris, N., Goldstein, J.M., Kennedy, D., Hodge, S.M., Caviness, V.S., Faraone, S.V., et al., 2006. Decreased volume of left and total anterior insular lobule in schizophrenia. Schizophr. Res. 83 (2), 155–171. https://doi.org/10.1016/j.schres.2005.11.020.

Marshall, I., Simonotto, E., Deary, I.J., Maclullich, A., Ebmeier, K.P., Rose, E.J., et al., 2004. Repeatability of motor and working-memory tasks in healthy older volunteers: assessment at functional MR Imaging. Radiology 233 (3), 868–877. https://doi.org/10.1148/radiol.2333031782.

McClure, S.M., Laibson, D.I., Loewenstein, G., Cohen, J.D., 2004. Separate neural systems value immediate and delayed monetary rewards. Science 306 (5695), 503–507. https://doi.org/10.1126/science.1100907.

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychol. Methods 1 (1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30.

Mejia, A.F., Nebel, M.B., Barber, A.D., Choe, A.S., Pekar, J.J., Caffo, B.S., Lindquist, M.A., 2018. Improved estimation of subject-level functional connectivity using full and partial correlation with empirical Bayes shrinkage. Neuroimage 172, 478–491. https://doi.org/10.1016/j.neuroimage.2018.01.029.

Mueller, S., Wang, D., Fox, M.D., Pan, R., Lu, J., Li, K., et al., 2015. Reliability correction for functional connectivity: theory and implementation. Hum. Brain Mapp. 36 (11), 4664–4680. https://doi.org/10.1002/hbm.22947.

Müller, K.U., Gan, G., Banaschewski, T., Barker, G.J., Bokde, A.L.W., Büchel, C., et al., 2015. No differences in ventral striatum responsivity between adolescents with a positive family history of alcoholism and controls: MID and family history alcohol. Addict. Biol. 20 (3), 534–545. https://doi.org/10.1111/adb.12136.

Mumford, J.A., Davis, T., Poldrack, R.A., 2014. The impact of study design on pattern estimation for single-trial multivariate pattern analysis. Neuroimage 103, 130–138. https://doi.org/10.1016/j.neuroimage.2014.09.026.

Nebe, S., Kroemer, N.B., Schad, D.J., Bernhardt, N., Sebold, M., Müller, D.K., et al., 2018. No association of goal-directed and habitual control with alcohol consumption in young adults: alcohol use and learning. Addict. Biol. 23 (1), 379–393. https://doi.org/10.1111/adb.12490.

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

Paul, E.J., Turner, B., Miller, M.B., Barbey, A.K., 2017. How Sample Size Influences the Reproducibility of Task-Based fMRI. 136259. https://doi.org/10.1101/136259.

Peters, J., Buchel, C., 2009. Overlapping and distinct neural systems code for subjective value during intertemporal and risky decision making. J. Neurosci. 29 (50), 15727–15734. https://doi.org/10.1523/JNEUROSCI.3489-09.2009.

Plichta, M.M., Schwarz, A.J., Grimm, O., Morgen, K., Mier, D., Haddad, L., et al., 2012. Test–retest reliability of evoked BOLD signals from a cognitive–emotive fMRI test battery. Neuroimage 60 (3), 1746–1758. https://doi.org/10.1016/j.neuroimage.2012.01.129.

Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R.J.A., Kahn, R.S., Ramsey, N.F., 2007. Test–retest reliability of fMRI activation during prosaccades and antisaccades. Neuroimage 36 (3), 532–542. https://doi.org/10.1016/j.neuroimage.2007.03.061.

Ripke, S., Hübner, T., Mennigen, E., Müller, K.U., Li, S.-C., Smolka, M.N., 2014. Common neural correlates of intertemporal choices and intelligence in adolescents. J. Cogn. Neurosci. 27 (2), 387–399. https://doi.org/10.1162/jocn_a_00698.

Ripke, S., Hübner, T., Mennigen, E., Müller, K.U., Rodehacke, S., Schmidt, D., et al., 2012. Reward processing and intertemporal decision making in adults and adolescents: the role of impulsivity and decision consistency. Brain Res. 1478, 36–47. https://doi.org/10.1016/j.brainres.2012.08.034.

Rodehacke, S., Mennigen, E., Müller, K.U., Ripke, S., Jacob, M.J., Hübner, T., et al., 2014. Interindividual differences in mid-adolescents in error monitoring and post-error adjustment. PLoS One 9 (2), e88957. https://doi.org/10.1371/journal.pone.0088957.

Rombouts, S.A.R.B., Barkhof, F., Hoogenraad, F.G.C., Sprenger, M., Scheltens, P., 1998. Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. Magn. Reson. Imag. 16 (2), 105–113. https://doi.org/10.1016/S0730-725X(97)00253-1.

Saddoris, M.P., Sugam, J.A., Stuber, G.D., Witten, I.B., Deisseroth, K., Carelli, R.M., 2015. Mesolimbic dopamine dynamically tracks, and is causally linked to, discrete aspects of value-based decision making. Biol. Psychiatry 77 (10), 903–911. https://doi.org/10.1016/j.biopsych.2014.10.024.

Safrit, M.J., 1976. Reliability theory. American Alliance for Health, Physical Education, and Recreation.

Schumann, G., Loth, E., Banaschewski, T., Barbot, A., Barker, G., Büchel, C., et al., 2010. The IMAGEN study: reinforcement-related behaviour in normal brain function and psychopathology. Mol. Psychiatr. 15 (12), 1128–1139. https://doi.org/10.1038/mp.2010.4.

Shou, H., Eloyan, A., Lee, S., Zipunnikov, V., Crainiceanu, A.N., Nebel, N.B., et al., 2013. Quantifying the reliability of image replication studies: the image intraclass correlation coefficient (I2C2). Cognit. Affect Behav. Neurosci. 13 (4), 714–724. https://doi.org/10.3758/s13415-013-0196-0.

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.

Spearman, C., 1910. Correlation calculated from faulty data. Br. J. Psychol. 3 (3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x, 1904-1920.

Story, G., Vlaev, I., Seymour, B., Darzi, A., Dolan, R., 2014. Does temporal discounting explain unhealthy behavior? A systematic review and reinforcement learning perspective. Front. Behav. Neurosci. 8. https://doi.org/10.3389/fnbeh.2014.00076.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., et al., 2015. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12 (3) e1001779. https://doi.org/10.1371/journal.pmed.1001779.

Taylor, R., 1990. Interpretation of the correlation coefficient: a basic review. J. Diagn. Med. Sonogr. 6 (1), 35–39. https://doi.org/10.1177/875647939000600106.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al., 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 den Bulk, B.G., Koolschijn, P.C.M.P., Meens, P.H.F., van Lang, N.D.J., van der Wee, N.J.A., Rombouts, S.A.R.B., et al., 2013. How stable is activation in the amygdala and prefrontal cortex in adolescence? A study of emotional face processing across three measurements. Dev. Cognit. Neurosci. 4, 65–76. https://doi.org/10.1016/j.dcn.2012.09.005.

Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T.E.J., Bucholz, R., et al., 2012. The Human Connectome Project: a data acquisition perspective. Neuroimage 62 (4), 2222–2231. https://doi.org/10.1016/j.neuroimage.2012.02.018.

Van Horn, J.D., Grafton, S.T., Miller, M.B., 2008. Individual variability in brain activity: a nuisance or an opportunity? Brain Imag. Behav. 2 (4), 327–334. https://doi.org/10.1007/s11682-008-9049-9.

Vetter, N.C., Pilhatsch, M., Weigelt, S., Ripke, S., Smolka, M.N., 2015. Mid-adolescent neurocognitive development of ignoring and attending emotional stimuli. Dev. Cognit. Neurosci. 14, 23–31. https://doi.org/10.1016/j.dcn.2015.05.001.

Vetter, N.C., Steding, J., Jurk, S., Ripke, S., Mennigen, E., Smolka, M.N., 2017. Reliability in adolescent fMRI within two years – a comparison of three tasks. Sci. Rep. 7 (1), 2287. https://doi.org/10.1038/s41598-017-02334-7.

Vul, E., Harris, C., Winkielman, P., Pashler, H., 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspect. Psychol. Sci. 4 (3), 274–290. https://doi.org/10.1111/j.1745-6924.2009.01125.x.

Waldvogel, D., van Gelderen, P., Immisch, I., Pfeiffer, C., Hallett, M., 2000. The variability of serial fMRI data: correlation between a visual and a motor task. Neuroreport 11 (17), 3843–3847.

Whelan, R., Watts, R., Orr, C.A., Althoff, R.R., Artiges, E., Banaschewski, T., et al., 2014. Neuropsychosocial profiles of current and future adolescent alcohol misusers. Nature 512 (7513), 185–189. https://doi.org/10.1038/nature13402.

Whitfield-Gabrieli, S., Nieto-Castanon, A., 2012. Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. Brain Connect. 2 (3), 125–141. https://doi.org/10.1089/brain.2012.0073.

Wilson, R.C., Niv, Y., 2015. Is model fitting necessary for model-based fMRI? PLoS Comput. Biol. 11 (6), e1004237.

Wittmann, M., Lovero, K.L., Lane, S.D., Paulus, M.P., 2010. Now or later? Striatum and insula activation to immediate versus delayed rewards. J. Neurosci., Psychol. Econom. 3 (1), 15–26. https://doi.org/10.1037/a0017252.

Yarkoni, T., Braver, T.S., 2010. Cognitive neuroscience approaches to individual differences in working memory and executive control: conceptual and methodological Issues. In: Handbook of Individual Differences in Cognition. Springer, New York, NY, pp. 87–107. https://doi.org/10.1007/978-1-4419-1210-7_6.

Yeo, B.T.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., et al., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol. 106 (3), 1125–1165. https://doi.org/10.1152/jn.00338.2011.

Zandbelt, B.B., Gladwin, T.E., Raemaekers, M., van Buuren, M., Neggers, S.F., Kahn, R.S., et al., 2008. Within-subject variation in BOLD-fMRI signal changes across repeated measurements: quantification and implications for sample size. Neuroimage 42 (1), 196–206. https://doi.org/10.1016/j.neuroimage.2008.04.183.