PLOS One

RESEARCH ARTICLE

Test–retest reliability of meta analytic networks during naturalistic viewing

Jean-Philippe Kröll[1,2]*, **Patrick Friedrich**[1,2], **Xuan Li**[1,2], **Yulia Nurislamova**[1,2], **Nevena Kraljevic**[1,2], **Anna Geiger**[1,2], **Julia Mans**[1,2], **Laura Waite**[1], **Julian Caspers**[3], **Xing Qian**[4], **Michael W. L. Chee**[5,6], **Juan Helen Zhou**[4,5,6], **Simon Eickhoff**[1,2], **Susanne Weis**[1,2]

1 Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich, Germany, **2** Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany, **3** Department of Diagnostic and Interventional Radiology, Medical Faculty and University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf, Germany, **4** Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore, **5** Centre for Sleep and Cognition & Centre for Translational MR Research, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore, **6** Integrative Sciences and Engineering Programme (ISEP), National University of Singapore, Singapore, Singapore

* j.kroell@fz-juelich.de

Check for updates

OPEN ACCESS

Citation: Kröll J-P, Friedrich P, Li X, Nurislamova Y, Kraljevic N, Geiger A, et al. (2026) Test–retest reliability of meta analytic networks during naturalistic viewing. PLoS One 21(5): e0346967. https://doi.org/10.1371/journal.pone.0346967

Editor: Akitoshi Ogawa, Sapporo Gakuin University, JAPAN

Received: April 11, 2025

Accepted: March 26, 2026

Published: May 6, 2026

Copyright: © 2026 Kröll et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: The Informed Consent Form of both datasets states that data is only available upon request of interested research groups and in anonymized form. For the JUMAX dataset, requests can be sent to the director of studies, PD Dr. Susanne Weis, Institut für Systemische Neurowissenschaften,

Abstract

Functional connectivity analyses have given considerable insights into human brain function and organization. As research moves towards clinical application, test-retest reliability has become a main focus of the field. So far, the majority of studies have relied on resting-state paradigms to examine brain connectivity, based on its low demand and ease of implementation. However, the reliability of resting-state measures is mostly poor to fair, potentially due to its unconstrained nature. Recently, naturalistic viewing paradigms have gained popularity because they probe the human brain under more ecologically valid conditions, possibly increasing reliability. We here compared the reliability of graph metrics extracted from resting-state and naturalistic viewing in functional networks, across two sessions. We show that naturalistic viewing can increase reliability over resting-state, but that its effect varies between stimuli and networks. Furthermore, we demonstrate that the effect of naturalistic viewing differs between two cohorts with Asian and European cultural backgrounds. Taken together, our study encourages the use of naturalistic viewing to increase reliability, but emphasizes the need to carefully select the appropriate stimulus for the network at hand.

1 Introduction

Functional magnetic resonance imaging (fMRI) data has become a widely-used tool to investigate neurological diseases and their underlying patterns [1–4]. The critical assumption behind all of these studies is that the measured brain activity is reliable, such that differences between subjects and timepoints are interpretable. However,

Heinrich Heine Universität Düsseldorf. Tel. 02461-61-8609, E-Mail: S.Weis@fz-juelich.de. Alternatively, data can be requested from the ethics committee of the Heinrich Heine University Düsseldorf, email: ethikkommission@med.uni-duesseldorf.de, tel: 0211/8119591.Data are stored on secure institutional servers with automated backups and redundancy to ensure long-term integrity. Data are organized in standardized formats with full metadata and version-controlled using DataLad. For the IMAX dataset, requests can be sent to the director of studies, Prof. Dr. Helen Juan Zhou, Centre for Translational Magnetic Resonance Research, Yong Loo Lin School of Medicine, National University of Singapore, E-Mail: helen.zhou@nus.edu.sg. Alternatively, the data can be requested from the National University of Singapore - Institutional Review Board, email: irb@nus.edu.sg, tel: +65 6516 4311. Data are stored on secure institutional servers and in standardized formats.

Funding: This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no. 945539 (HBP SGA3), and the Deutsche Forschungsgemeinschaft (491111487). We also acknowledge the funding support from Yong Loo Lin School of Medicine, National University of Singapore (J.H.Z), the Duke-NUS Medical School Signature Research Program Core Funding (J.H.Z.), and Ministry of Education, Singapore (MOE-T2EP40120-0007, J.H.Z), and Far East Organization (E-546-00-0398-01, MWLC.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors declare that they have no conflict of interest.

the reported reliability of fMRI measures varies vastly across studies [5], due to small test-retest samples and different analysis choices. As fMRI research moves towards the identification of biomarkers [3,6–8], increasing reliability has become a priority. In order to aid in the diagnosis and prognosis of brain disorders, a measure has to be capable of giving consistent results, otherwise it is unsuitable as a biomarker.

The majority of prior reliability studies have relied on metrics derived from resting state (RS) [9–11]. With low demands on the participants, RS is well suited for healthy as well as patient cohorts and allows for a quick data acquisition. The RS paradigm also suffers from a few drawbacks. Data acquired during RS can be strongly confounded by head movement and drowsiness of the participant due to its unconstrained nature [12,13], as participants struggle to remain awake and motionless in the absence of a task or stimulus. For the same reasons, RS is more susceptible to be influenced by spontaneous thought of the participant [14,15]. Therefore, it comes as no surprise that reliability of RS functional connectivity (FC) is generally considered poor to fair [16,17].

Naturalistic Viewing (NV) paradigms, during which participants are presented with a story or a film, have recently gained popularity because they might give insight into the brain's function under more ecologically valid conditions. It has been shown that NV poses several advantages over conventional RS such as increased participant engagement, reduced head movement and increased synchronization between subjects [18,19]. Especially relevant for clinical studies, NV shares with RS the advantage of minimizing demand on the participants [20]. On the other hand, NV paradigms place a behavioral constraint that allows for the study of normal and abnormal brain function, somewhat similar to task-based designs. Making use of these advantages, a series of studies could show altered connectivity during NV in patients [21–24], encouraging the application of NV measures as biomarkers.

Furthermore, several studies suggest that NV increases test-retest reliability in comparison with RS [19,25,26]. This improvement can be attributed to several factors. First, many studies have pointed out that NV improves signal properties by increasing participant engagement [20,27–29]. Secondly, by reducing head movement and drowsiness, NV is less susceptible to noise than conventional RS. Thirdly, by presenting the same stimulus across sessions, NV is less influenced by spontaneous thought of the participant while also placing a behavioral constraint that reduces variance. However, the effect of NV on reliability is dependent on various factors such as attention [30], successful episodic encoding [31] as well as the chosen movie stimulus [32–34] and differs between different brain regions and networks [19].

Based on findings showing that human brain organization shares features of complex networks, such as being organized according to small-world properties [35,36], many studies have used graph theoretical approaches to analyze fMRI data [9,37,38]. Different studies reported low to good reliability when investigating the reliability of graph measures acquired during RS. Braun et al. found moderate reliability for the graph metrics clustering coefficient, characteristic path length, local and global efficiency, assortativity, modularity, hierarchy and the small-worldness

scalar [9]. Deuker et al reported good reliability (mean ICC = 0.62) for a similar set of graph measures extracted from MEG data recorded during the n-back task, but observed decreased reliability during RS [38]. Wang 2011 reported mostly low reliability for graph metrics extracted from RS data [11].

The application of graph theory to fMRI data offers additional insights over FC analyses by providing quantitative metrics that characterize the organization and efficiency of brain networks [39]. In addition, several studies could show that graph metrics are well suited to capture network connectivity abnormalities caused by neurological diseases, therefore making the use of graph theory attractive to clinical research [3,6,7].

A first study by Wang et al compared the test-retest reliability of graph measures extracted from RS and one movie condition in a sample of 17 participants [19]. They found that reliability of five different graph measures was generally increased during NV in comparison with RS. However, the authors themselves acknowledge that their study does not take into account how the choice of the movie or differences between populations might impact the measure. Previous studies on FC during NV showed that different movie stimuli vary in their ability to predict behavior [27], increase within- and between-subject correlations [29] and improve identifiability [29,34]. Therefore, it is likely that different movie stimuli also impact the reliability of graph measures extracted from NV. In addition, several authors have suggested that the same NV stimuli might deviate in their effect between different populations [20,32,40]. The cultural background of a participant is likely to influence how a given movie is perceived and might result in deviating effects across cohorts.

The present study aims to further evaluate the test-retest reliability of NV, by investigating its influence on the reliability of five commonly used graph theoretical measures. To benchmark the reliability of NV, we compare it to that of RS. Further, we evaluate the influence of the movie content, by employing stimuli with different levels of social content, ranging from the neutral movie *Inscapes*, over the silent movie *The Circus*, to the most social movie *Indiana Jones and the Temple of Doom*. In addition, to study the effect of NV in different populations, we here compare the results from two independent samples from Europe and Asia, respectively, using the same stimuli. In contrast to the majority of previous studies, we here compare reliability on the basis of a priori defined networks, and not on a whole-brain basis.

The analysis of network based measures allows us to investigate how NV influences the reliability in different cognitive domains. The networks implemented in this study are meta-analytically defined networks that represent the most likely core nodes involved in a given cognitive function, because they incorporate convergent information from a multitude of studies. The networks were chosen to cover a broad range of functional domains (e.g., affective, social, executive, memory and motor functions).

2 Methods

2.1 Participants

Two datasets were used in this study. The first dataset, IMAX, was collected at the Centre for Translational MR Research, National University of Singapore, and the second dataset, JUMAX, was collected at the Forschungszentrum Jülich. Exclusion criteria for both studies were neurological or psychiatric diagnoses, significant visual or hearing impairment, alcohol or caffeine consumption 6 hours prior to the scan and self-reporting of bad sleep the night before the scan days. All participants underwent three identical testing sessions within a one-week interval. Subjects gave written, informed consent and were compensated for their participation. The study was approved by the institutional review board of the National University of Singapore for IMAX and by the ethics committee of the Heinrich Heine University, Düsseldorf for JUMAX. Due to unavailability of part of the data of the JUMAX sample, the final cohort comprised 33 subjects (14 females, mean age 27.5 + /- 3 years). Accordingly, to match the number of available subjects from the JUMAX dataset, only the first 33 subjects were used from the IMAX sample (17 females, 27 + /- 2.7) (Fig 1).

To compare head motion between the two samples, an ANOVA was performed that revealed no statistically significant difference in framewise displacement between the samples, conditions or sessions (S1 Table).

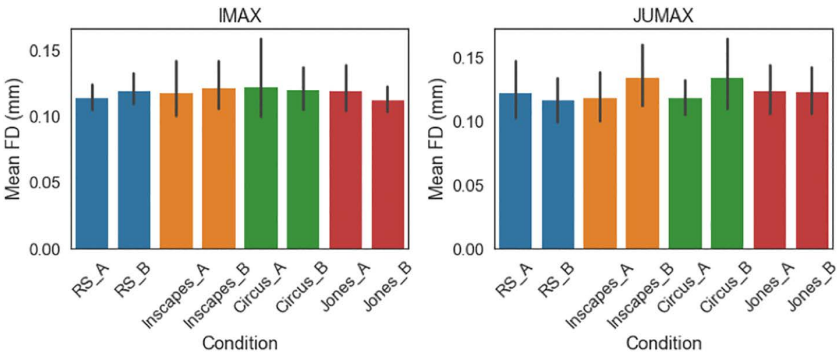

Fig 1. Head motion comparison. Framewise Displacement (FD) under the different conditions (A = Session A, B = Session B, RS = resting state, Jones = Indiana Jones).

https://doi.org/10.1371/journal.pone.0346967.g001

2.2 Data acquisition

For both datasets, the data was acquired on a Siemens Magnetom PrismaFit 3-Tesla with a 20-Channel head coil. Structural images were collected using an MP-RAGE sequence (TR = 2300ms, TE = 2,28ms, TI = 900ms, flip-angle = 8°) and 1 mm voxel size. All RS and NV runs used the same echo planar imaging sequence (TR = 719ms, TE = 30ms, flip-angle = 52°, slices = 44, FOV = 225x225 mm2) resulting in 2.96x2.96x3 mm voxel size. Data from collaborators at the National University of Singapore were retrieved and structured in the form of a DataLad dataset, a research data management solution providing data versioning, data transport, and provenance capture [41]. Each of the three testing sessions per participant, which were conducted within a seven day period, comprised three NV runs and two RS scans. The order of scans was identical on all three days, starting with a structural scan, followed by 5 functional scans in the order of RS 1, Inscapes, Circus, Indiana Jones and RS 2, with each functional scan lasting for 10 minutes. All movies had been cut to the same length. For RS scans, participants were asked to lay as still as possible and think of nothing in particular, while keeping their eyes open. Instructions for the NV scans were to watch the movies while staying as still as possible. For all scans, participants were asked to not fall asleep during the measurement. Foam wedges were fitted around each subject's head for comfort and to decrease movement. For all analyses, only the first two sessions of both samples were used to ensure comparability with previous literature and avoid further influence due to repeated viewing. Furthermore, only the first RS scan was considered to avoid possible effects on RS from prior movie watching [42]. The movie clips were presented via a mirror that was mounted on the head coil and the sound was played through headphones.

2.3 Stimulus material

Three different movie stimuli with different levels of social content (Inscapes < The Circus < Indiana Jones) were used. Inscapes is a nonverbal, non-social series of animated abstract shapes created by Vanderwal et al. which was looped to match the 10 minutes duration [43]. The Circus (United Artists Digital Studios, 1928, directed by Charlie Chaplin) is a silent black-and-white. Participants were shown the first 10 minutes of the film which depicts the protagonist being chased by the police and unintentionally causing comic situations during his escape. Indiana Jones and the Temple of Doom (Paramount Pictures, 1984, directed by Steven Spielberg) shows the first 10 minutes of the movie during which the protagonist has to fight off several hitmen who are trying to kill him and finally escapes by taking a plane. The end of the clips used from The Circus and Indiana Jones both coincide with a change of scene in the respective movie itself.

2.4 Data preprocessing

Preprocessing of MRI data was performed using fMRIPrep, version 22.0.0 [44]. In brief, the T1-weighted volumes were corrected for intensity non-uniformity and skull-stripped. The extracted brain images were then transformed into Montreal Neurological Institute (MNI) space and motion corrected using Advanced Normalization Tools [45]. The functional data was motion-corrected with MCflirt [46] and subsequently co-registered to the native T1-weighted image using boundary based registration with six degrees of freedom from Freesurfer [47]. Subsequently, an isotropic Gaussian kernel of 6 mm FWHM (full-width half-maximum) was applied for spatial smoothing. The images were further regressed out of nuisance signals and bandpass filtered (0.01–0.1 Hz). Nuisance signals were the global signals extracted within the CSF, the WM, and the whole-brain masks which were regressed from the preprocessed fMRI data for each subject. In addition, the standard six motion parameters and their first temporal derivatives were regressed out.

Subsequently, network functional connectivity (NFC) matrices were constructed for 14 meta-analytical networks, comprising nine to 23 nodes (a detailed description of the networks can be found in the supplements). In short, isotropic 5 mm spheres were created around the local maxima of each meta-analytical network node and only gray matter voxels were included. Using the Junifer toolbox [48], we extracted the mean time series of each node and computed the covariance between all node pairs to produce a node times node connectivity matrix for each subject and each condition. The networks cover affective [49–52], social [49,53,54], executive [55–58], memory [59,60] and motor [61] functions.

2.5 Graph theoretical analyses

Subsequently, graph metrics were derived from the NFC matrices. The fully connected node x node matrices were thresholded at 0.1 to determine the presence or absence of connections (edges) between nodes. Connections above the threshold retained their correlation coefficient, whereas subthreshold edges were assigned values of 0. This thresholding procedure was performed on both positive and negative connections.

Five different Graph metrics were extracted from the thresholded NFC matrices using the *NetworkX* toolbox [62], including degree centrality, clustering coefficient, betweenness centrality, global efficiency and mean shortest path length. Degree centrality measures the connectedness of each node, computed as the weighted sum of all edges connected to that node. The clustering coefficient for a given node is a measure of local connectedness, measuring the proportion of existing connections out of all possible connections between the nearest neighbors of that node. Betweenness centrality measures the centrality of a node in the network, calculated as the ratio of shortest paths (that is the smallest number of links that need to be traversed to go from one node to another) in the whole graph that pass through that node. The efficiency of a pair of nodes in a graph is the reciprocal of the shortest path distance between these two nodes. The global efficiency of a graph is the average efficiency of all pairs of nodes. Shortest path length denotes the minimum number of nodes that need to be passed through to connect one node to another. Mean shortest path length is the average shortest path length between all nodes of the graph.

2.6 Test-retest reliability

The reliability of each graph metric was quantified by calculating the intraclass correlation coefficient (ICC) across these measures derived from the two scans [63,64]. A two-way mixed-effects ANOVA model (subjects × scan session) was applied to the measures derived from the two scan sessions across subjects to obtain the between-subject mean square (MSp) and mean square error (MSe). ICC values were then calculated as:

$$ICC\,(3,1) \;=\; MSp \;-\; MSe \,/\, MSp + (d-1)\,MSe$$

where d is equal to the number of observations per subject. For each graph measure, we calculated reliability at the scan-wise level. Scan-wise reliability estimates the reliability of one score derived from the entire scan session, opposed to

calculating one ICC value for the graph metric of each node [10,19]. Here, a single ICC value was calculated for the mean graph metric averaged across all nodes of the network. The reliability results are considered excellent (ICC > 0.8), good (ICC 0.6–0.79), moderate (ICC 0.4–0.59), fair (ICC 0.2–0.39), and poor (ICC < 0.2) [65]. As negative ICCs are difficult to interpret and reasons for negative values are unclear [66], in the following we set negative ICCs to zero (that is completely non-reliable) as has been suggested in previous studies [9,67,68].

3 Results

3.1 Reliability of graph metrics in the IMAX sample

We investigated the reliability of five graph measures derived from 14 different networks. For the IMAX sample, we found low to good reliability across networks. Degree centrality, cluster coefficient and efficiency showed a trend towards higher reliability than between centrality and shortest path length (Fig 2).

Degree centrality showed the highest ICC during RS in five (AM, CogAC, MNS, Rew and VigAtt), during Inscapes in three (SM, ER, extDMN), during Circus in four (EmoSF, Empathy, ToM, WM) and during Jones in three (eSAD, Motor, Empathy) networks.

Cluster coefficient showed the highest ICC during RS in four (Rew, Empathy, VigAtt, EmoSF), during Inscapes in three (MNS,ER,extDMN), during Circus in three (AM, SM, ToM) and during Jones in five (CogAC, Motor, EmoSF, eSAD, WM) networks.

Efficiency showed the highest ICC during RS in eight (AM, MNS, CogAC, EmoSF, Rew, eSAD, extDMN, WM), during Inscapes in three (Motor, SM, ER, extDMN), during Circus in three (Empathy, ToM, WM) and during Jones in two (VigAtt, WM) networks.

Between centrality showed the highest ICC during RS in two (EmoSF, Empathy), during Inscapes in five networks (AM, MNS, SM, eSAD, extDMN), during Circus in four (Motor, ER, ToM, WM) and during Jones in three (CogAC, Rew, VigAtt) networks.

Shortest path length showed the highest ICC during RS in four (MNS, EmoSF, eSAD, WM), during Inscapes in four (AM, Motor, Empathy, extDMN), during Circus in two (ER, ToM) and during Jones in four (CogAC, Rew, SM, VigAtt) networks.

3.2 Reliability of graph metrics in the JUMAX sample

For JUMAX we found low to excellent reliability across networks. Degree centrality, cluster coefficient and efficiency showed a trend towards higher reliability than between centrality and shortest path length (Fig 3).

Degree centrality showed the highest ICC during RS in nine (AM, CogAC, EmoSF, Empathy, ER, MNS, Motor, VigAtt, WM), during Inscapes in one (eSAD) and during Circus in three (Rew, SM, ToM) networks.

Cluster coefficient showed the highest ICC during RS in five (CogAC, Motor, EmoSF, SM, WM), during Inscapes in two (AM, eSAD), during Circus in four (Rew, ER, ToM, VigAtt) and during Jones in three (MNS, Empathy, extDMN) networks.

Efficiency showed the highest ICC during RS in ten (AM, MNS, CogAC, Motor, EmoSF, Empathy, ER, eSAD, VigAtt, extDMN), during Circus in three (Rew, SM, WM) and during Jones in one (ToM) networks.

Between centrality showed the highest ICC during RS in four (AM, Motor, Rew, ER), during Inscapes in two (Empathy, ToM), during Circus in one (SM) and during Jones in seven (CogAC, EmoSF, MNS, eSAD, VigAtt, extDMN, WM) networks.

Shortest path length showed the highest ICC during RS in nine (AM, MNS, CogAC, EmoSF, Rew, ER, SM, ToM, VigAtt, extDMN), during Circus in two (Empathy, SM) and during Jones in three (Motor, eSAD, WM) networks.

3.3 Comparison of the two samples

Comparing the results across the two samples, it was evident that the ICC was generally higher in the JUMAX sample than in the IMAX sample. However, in both samples, degree centrality and efficiency tended to show the highest ICCs,

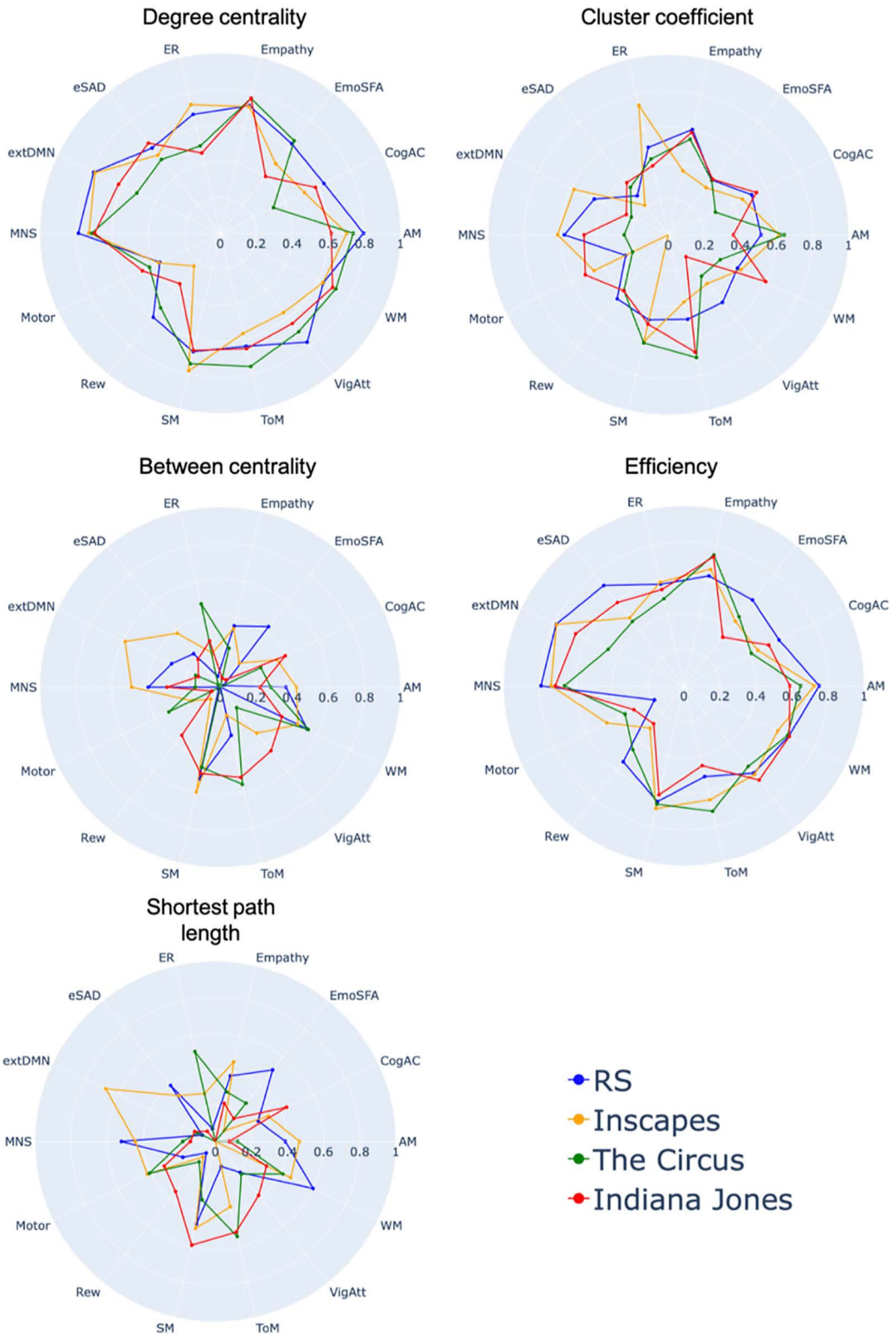

Fig 2. ICC of graph metrics across the 14 networks in the IMAX sample. Graph metrics are shown for the RS scan and three different movies. ICC values below zero are not depicted. (AM = Autobiographical Memory, CogAC = Cognitive Attention Control,eMDN = extended Multiple Demand Network,

EmoSF = Emotional Scene and Face Processing, ER = Emotion Regulation, eSAD = Extended Social-affective Default, MNS = Mirror Neuron System, Rew = Reward, SM = Semantic Memory, ToM = Theory of Mind, VigAtt = Vigilant Attention, WM = Working memory).

https://doi.org/10.1371/journal.pone.0346967.g002

followed by cluster coefficient and then by between centrality and shortest path length. The AM, MNS, Empathy and SM networks showed similar results in both samples, while the rest of the networks showed more distinct results. Overall there was not one stimulus which led to more consistent results than other stimuli across the two samples (Fig 4).

4 Discussion

The primary goal of this study was to investigate the reliability of NV and RS, across various functional networks. Graph metrics indicate that NV is – in certain conditions – more reliable than RS, consistent with previous results from Wang et al. 2017 [19]. However, our results demonstrated that this effect is dependent on a variety of factors. Firstly, the choice of the NV stimulus impacts the reliability of a given graph metric. Secondly, the effect of NV stimuli varies across cohorts. Thirdly, the increase in reliability is not uniform across the brain, but varies between different functional networks. Different graph metrics can be taken to have different implications with respect to functional brain networks. For example, shortest path length and efficiency have been interpreted as measures of how efficiently information is transferred in the brain [35]. Degree- and between-centrality are commonly seen as measures of the importance of single nodes in a network and can help to identify critical hubs [37]. Clustering coefficient reflects the interconnectedness of nodes and is often interpreted as an indicator for specialized brain modules [69]. However, in the following we will summarize the main findings instead of focusing on single graph metrics. Although each of these measures capture unique aspects of brain organization, the focus of this paper is on the reliability of NV vs RS, using graph metrics as a proxy.

4.1 NV vs RS

Starting from observations indicating that graph metrics extracted from RS fMRI can be used to investigate abnormalities in brain organization [4,39], researchers have focused on investigating the reliability with which these graph metrics can be extracted. With ongoing efforts to use characteristic abnormalities to successfully detect and track neurological diseases, it will be crucial to increase reliability as much as possible.

Therefore, researchers have shifted to extracting graph metrics from other modalities than RS such as task-based fMRI [70,71] or NV [72,73]. In contrast to task-free RS, these modalities place a constraint on the participant which might reduce variability that is otherwise induced by spontaneous thoughts [29,32,74]. Our results confirmed the notion that behavioral constraints can prove to be beneficial to increase reliability over unconstrained RS. In multiple networks, NV stimuli increased reliability of one or more graph metrics in comparison with RS, at least numerically. Furthermore, this improvement of reliability is observable across networks dealing with affective, social, executive, memory and motor functions, indicating that NV increases engagement not only in sensory, but also in higher order networks. On the other hand, our results also indicated that in many instances RS was more reliable than NV, which is in line with previous studies that showed that NV does not unconditionally increase reliability [26,75]. Nevertheless, these results, in our opinion, encourage the use of NV to improve reliability as NV increased reliability over RS drastically in certain cases. Rather than viewing NV as a one-fits-all tool, our findings further underline the importance of using specific NV stimuli (and brain networks) for a specific purpose.

The observed reliability in our study matches results from previous studies investigating graph metrics extracted from RS and NV [9,19,71]. However, in contrast to Wang et al [19] we showed that NV does not generally improve reliability of graph metrics, but that its effect varies across networks, stimuli and graph metric.

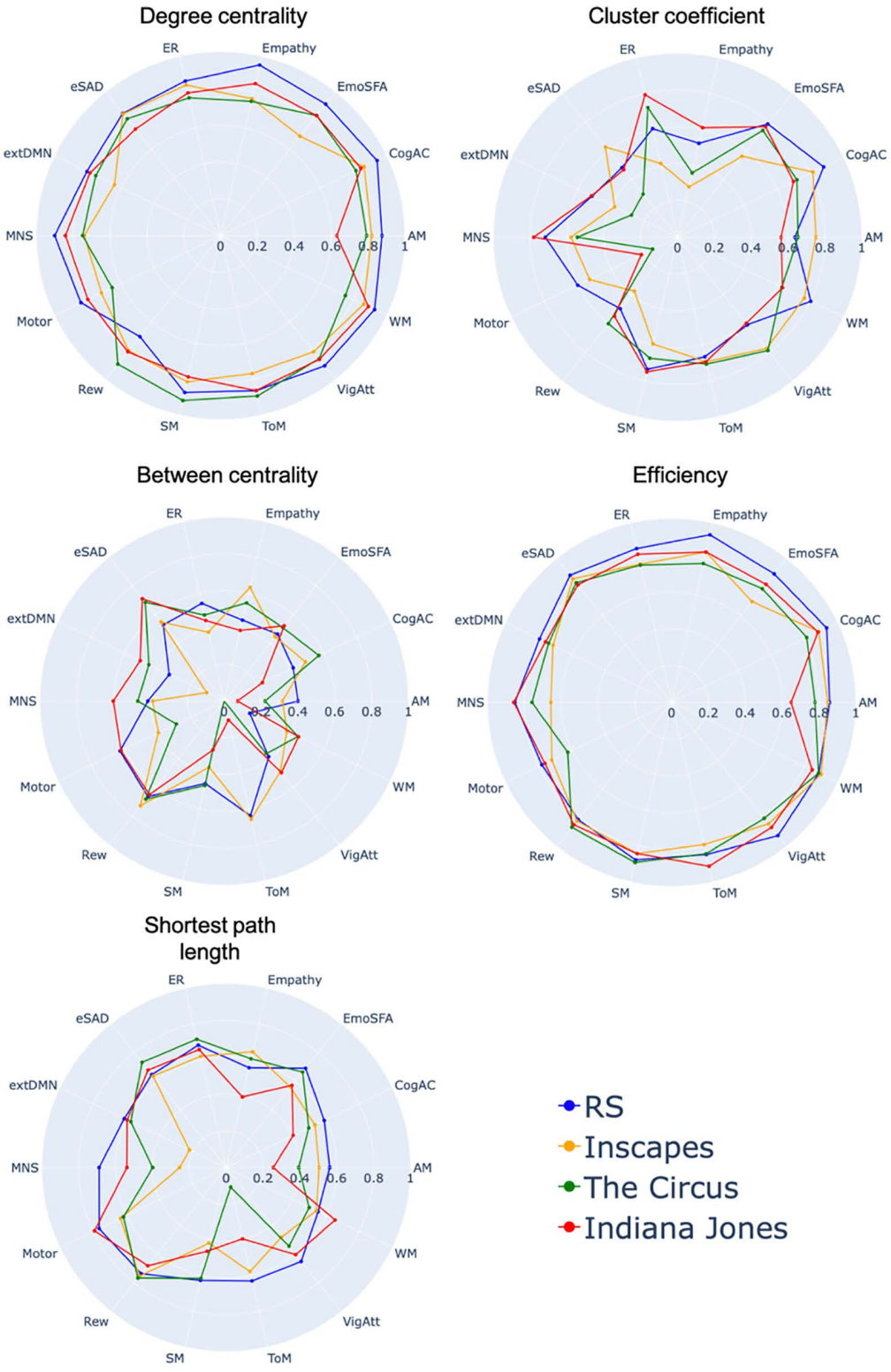

Fig 3. ICC of graph metrics across the 14 networks in the JUMAX sample. Graph metrics are shown for the RS scan and three different movies. ICC values below zero are not depicted. (AM = Autobiographical Memory, CogAC = Cognitive Attention Control,eMDN = extended Multiple Demand

Network, EmoSF = Emotional Scene and Face Processing, ER = Emotion Regulation, eSAD = Extended Social-affective Default, MNS = Mirror Neuron System, Rew = Reward, SM = Semantic Memory, ToM = Theory of Mind, VigAtt = Vigilant Attention, WM = Working memory).

https://doi.org/10.1371/journal.pone.0346967.g003

4.2 Variance across cohorts

One of the advantages of using NV stimuli is that they are easier to share across multiple sites than traditional tasks [20,76]. By combining data from a multitude of studies using the same NV stimuli, one can not only achieve large sample sizes, but also place the same behavioral constraint on all subjects across sites. However, several studies have suggested that cultural differences between movies (and/or cohorts) might hinder generalizability [20,32,76]. In this study, we compared an asian and a european cohort that were subjected to the same three NV stimuli. In both samples, NV stimuli appear to increase reliability of graph metrics in comparison with RS. However, we did not observe that the same combination of stimulus, network and graph metric led to improved reliability over RS in both samples. Although some of the networks (AM, MNS, Empathy and SM) show similar trends, it is not generally the case that results from both samples are highly overlapping. These differences might have been driven by the different cultural backgrounds of the participants. The appreciation of a film is culturally specific [77] and likely different between the european and asian cohorts. Several studies have demonstrated cultural differences in the perception of faces [78–80], a factor that is especially relevant for the NV stimuli Circus and Jones during which a variety of different faces are depicted. Related, in a study from Sneddon et al., 2011 [81] participants from Northern Ireland, Serbia, Guatemala and Peru showed systematic differences in their rating of positive and negative emotions being displayed in twelve short movie clips. Our study provides further evidence for the notion that future studies should take into account cultural differences between cohorts when selecting a movie stimulus.

4.3 Variance across networks and stimuli

In our analysis, we employed meta-analytically defined networks that represent the most likely core nodes of a given brain function. Alternatively to approaches where the effect of NV is considered from a whole brain perspective, we here investigated how NV engages different networks. Similar to previous studies, we observed that the effect of different NV stimuli varies across different networks [19,27,33] and reliability of graph metrics was not unconditionally increased over RS. One of the advantages of NV is the possibility to more effectively engage brain networks of interest, in comparison with RS [20,22]. Intuitively, one would expect that a network responsible for the processing of emotions is differently engaged by an emotional clip than, e.g., the motor network. This effect can also be seen in our results as different networks exhibit varying reliabilities in response to the same stimulus

To analyze the effect of the chosen movie stimulus on the reliability of a given graph metric, we employed three movies with different levels of social content. Various studies have shown that different NV stimuli can lead to significantly different results. Finn et al., 2021 reported that FC derived from different movies varied in its ability to accurately predict emotion and cognition scores [27]. Similarly, Gal et al., 2022 showed that the accuracy with which task-activation maps could be predicted differed between FC derived from Hollywood NV stimuli and independent NV stimuli [82]. Our results extend these findings by showing that NV stimuli also divert in their impact on the reliability of extracted graph measures. Previous studies have shown that reliability is strongly dependent on attention [30] and several studies have suggested that NV stimuli with social content are best suited to engage participants and keep their attention over a longer period of time [27,77,83]. In line with that, the NV effects varied across the three movie clips. Inscapes, which contains abstract visual patterns and minimal social content, showed reliability advantages primarily in attention- and memory-related networks (AM, extDMN, SM, but also ER). Circus, which includes moderate levels of social interaction, showed improvements mainly in the theory of mind network, but also in the SM in JUMAX. Jones, which contains the richest narrative and

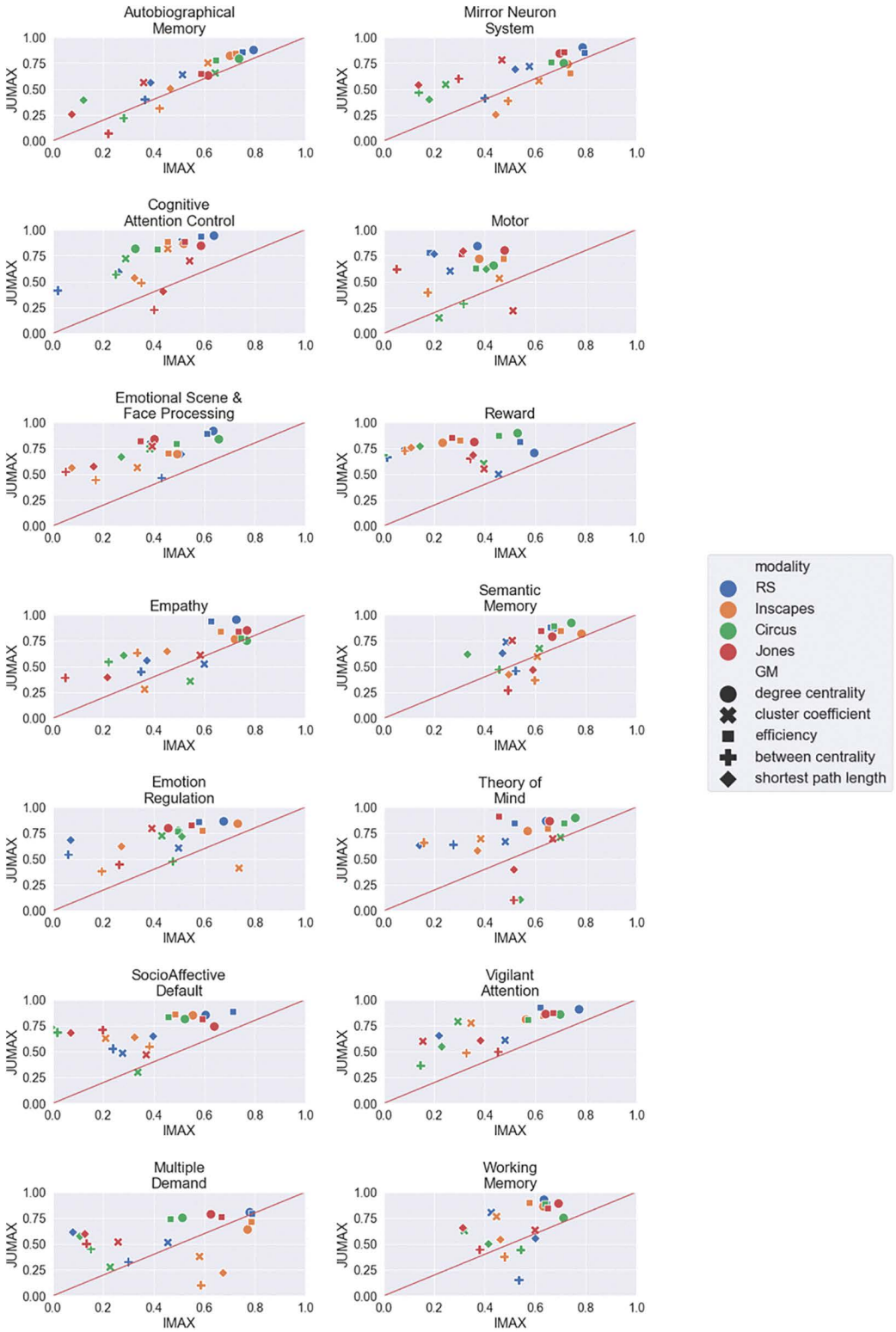

Fig 4. Comparison of ICC between IMAX and JUMAX. (RS = resting state, Jones = Indiana Jones).

https://doi.org/10.1371/journal.pone.0346967.g004

social content, most frequently produced the highest ICC values in CogAC and VigAtt. A full overview can be found in the supplements (S1 and S2 Tables). These results suggest that the relative reliability of NV compared to RS depends on the specific stimulus characteristics and the functional systems engaged by the stimulus.

However, in the majority of cases, reliability was higher for graph metrics extracted from RS than these extracted from NV. This was somewhat unexpected since RS is generally seen as an unconstrained state and one would therefore expect more variability between sessions than for more constrained states like NV. Several factors might have led to the relative decrease in reliability for NV. Firstly, familiarity with a given movie might have played a role as multiple studies have shown that expected stimuli reduce the neuronal response [84,85]. The sessions for both datasets were conducted within a week and therefore participants will be familiar with the movie during the second session. This effect might have induced variability for the NV conditions, while RS on the other hand has been shown to remain stable across sessions [11,86]. Secondly, some of the networks employed here (AM, SM and eSAD) are overlapping with the default mode network which is linked to intrinsically oriented functions, rather than the processing of external stimuli [18,87]. This may plausibly lead to decreased reliability of NV in comparison with RS, in these networks.

These results emphasize that future studies should carefully consider which combination of graph metric, stimulus and network is suited for the research question at hand. Using purpose-built movies, such as emotionally salient clips for patients with depression [21], in combination with the functionally involved network will help improve reliability and advance the characterization of disease specific alterations in the brain.

5 Limitations

While the current study sheds new light onto the reliability of NV in comparison with RS, it comes with some limitations. Firstly, the reliability of graph metrics is strongly influenced by the choice of the applied preprocessing [88]. In this study, we applied motion correction and regressed out WM and CSF signals, as has been done in most previous studies [9,19,71]. On top of that, we here applied basic (that is only removing the mean signal of the whole brain) global signal regression. There is an ongoing debate of whether or not to apply global signal regression, with some studies claiming that it introduces spurious anti-correlations while other reports suggest that these anti-correlations are true negative connections [89,90]. However, a review by Andellini et al., 2015 found no significant differences between the reliability of data with and without the inclusion of global signal regression across five studies [88]. Secondly, we here considered both, negative and positive connections, with the assumption that both are true representations of connectivity. However, several papers have indicated that negative correlations should be evaluated with care since they tend to reduce test-retest reliability [11,88,91]. Therefore, the reliability of single graph metrics in our study might have been decreased by the inclusion of negative connections. Thirdly, our results are based on weighted adjacency matrices, because they better characterize the underlying connectivity by considering connectivity strength. However, previous studies have suggested that binarized adjacency matrices may lead to higher reliability [11,88]. Nevertheless, we think that using weighted adjacency matrices is preferable, especially for clinical studies where subtle changes in connectivity might help to identify disease specific alterations.

In addition, the scan order might have influenced our results as RS was always acquired before NV and participants might have become more settled after the first scan. However, it has been shown that prior cognitive states may influence RS-FC and therefore this design is preferable to avoid effects on RS from prior movie watching [42]. On top of that, the sequence of the movies was consistent across subjects. A balanced scanning scheme might strengthen the comparison between NV stimuli. Finally, the current literature does not agree on a method for statistically comparing test-retest reliability metrics between groups or conditions and therefore we here relied on a more descriptive approach.

6 Conclusion

NV has been suggested to improve the reliability of graph based measures in comparison with RS. Our findings extend the current knowledge by investigating this effect in different networks, with multiple NV stimuli and in two different

cohorts. We demonstrate that the potential increase in reliability is dependent on the chosen NV stimuli and varies between functional networks. Furthermore we suggest that cultural differences should be considered when sharing NV stimuli across sites. Our study supports the use of NV to increase reliability of graph metrics, but emphasizes the need to carefully select the appropriate stimulus for the network of interest. Based on our results we generally suggest to implement stimuli with rich social content. Future studies might profit from using purpose-built movies, such as thrilling or sad stimuli to engage emotion networks or dialog heavy scenes to investigate language processing networks.

Supporting information

S1 Table. ANOVA revealed that there was no statistically significant difference in framewise displacement between the samples, conditions or sessions.
(DOCX)

S2 Table. Condition yielding the highest ICC for each functional network and graph metric in IMAX.
(DOCX)

S3 Table. Condition yielding the highest ICC for each functional network and graph metric in JUMAX.
(DOCX)

S4 Table. Centroid Coordinates of the nodes of the meta-analytically defined networks.
(DOCX)

S1 Fig. Permutation tests of the reliability difference of different graph measures across RS and NV conditions. ICCs are compared to corresponding null distribution with 5,000 randomizations. Vertical lines indicate the observed values in each condition. Blue lines indicate 95% CIs. The red vertical line indicates the observed difference. Only plots with significant results after FDR-BH correction are shown.
(DOCX)

S2 Fig. Node location of the meta-analytically defined networks.
(DOCX)

Author contributions

Conceptualization: Jean-Philippe Kröll, Susanne Weis.

Data curation: Jean-Philippe Kröll, Laura Waite, Xing Qian.

Funding acquisition: Simon Eickhoff.

Methodology: Susanne Weis.

Software: Jean-Philippe Kröll.

Supervision: Susanne Weis.

Writing – original draft: Jean-Philippe Kröll.

Writing – review & editing: Patrick Friedrich, Xuan Li, Yulia Nurislamova, Nevena Kraljevic, Anna Geiger, Julia Mans, Julian Caspers, Xing Qian, Michael WL Chee, Juan Helen Zhou, Simon Eickhoff, Susanne Weis.

References

1. Balthazar MLF, de Campos BM, Franco AR, Damasceno BP, Cendes F. Whole cortical and default mode network mean functional connectivity as potential biomarkers for mild Alzheimer's disease. Psychiatry Res. 2014;221(1):37–42. https://doi.org/10.1016/j.pscychresns.2013.10.010 PMID: 24268581

2. Basaia S, Agosta F, Wagner L, Canu E, Magnani G, Santangelo R, et al. Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. Neuroimage Clin. 2019;21:101645. https://doi.org/10.1016/j.nicl.2018.101645 PMID: 30584016
3. Supekar K, Menon V, Rubin D, Musen M, Greicius MD. Network analysis of intrinsic functional brain connectivity in Alzheimer's disease. PLoS Comput Biol. 2008;4(6):e1000100. https://doi.org/10.1371/journal.pcbi.1000100
4. Wu T, Wang L, Chen Y, Zhao C, Li K, Chan P. Changes of functional connectivity of the motor network in the resting state in Parkinson's disease. Neurosci Lett. 2009;460(1):6–10. https://doi.org/10.1016/j.neulet.2009.05.046 PMID: 19463891
5. Bennett CM, Miller MB. How reliable are the results from functional magnetic resonance imaging?. Ann N Y Acad Sci. 2010;1191:133–55. https://doi.org/10.1111/j.1749-6632.2010.05446.x PMID: 20392279
6. Bassett DS, Bullmore E, Verchinski BA, Mattay VS, Weinberger DR, Meyer-Lindenberg A. Hierarchical organization of human cortical networks in health and schizophrenia. J Neurosci. 2008;28(37):9239–48. https://doi.org/10.1523/JNEUROSCI.1929-08.2008 PMID: 18784304
7. Rubinov M, Knock SA, Stam CJ, Micheloyannis S, Harris AWF, Williams LM, et al. Small-world properties of nonlinear brain activity in schizophrenia. Hum Brain Mapp. 2009;30(2):403–16. https://doi.org/10.1002/hbm.20517 PMID: 18072237
8. Wang L, Zhu C, He Y, Zang Y, Cao Q, Zhang H, et al. Altered small-world brain functional networks in children with attention-deficit/hyperactivity disorder. Hum Brain Mapp. 2009;30(2):638–49. https://doi.org/10.1002/hbm.20530 PMID: 18219621
9. Braun U, Plichta MM, Esslinger C, Sauer C, Haddad L, Grimm O, et al. Test-retest reliability of resting-state connectivity network characteristics using fMRI and graph theoretical measures. Neuroimage. 2012;59(2):1404–12. https://doi.org/10.1016/j.neuroimage.2011.08.044 PMID: 21888983
10. Guo CC, Kurth F, Zhou J, Mayer EA, Eickhoff SB, Kramer JH, et al. One-year test-retest reliability of intrinsic connectivity network fMRI in older adults. Neuroimage. 2012;61(4):1471–83. https://doi.org/10.1016/j.neuroimage.2012.03.027 PMID: 22446491
11. Wang J-H, Zuo X-N, Gohel S, Milham MP, Biswal BB, He Y. Graph theoretical analysis of functional brain networks: test-retest evaluation on short- and long-term resting-state functional MRI data. PLoS One. 2011;6(7):e21976. https://doi.org/10.1371/journal.pone.0021976 PMID: 21818285
12. Tagliazucchi E, Laufs H. Decoding wakefulness levels from typical fMRI resting-state data reveals reliable drifts between wakefulness and sleep. Neuron. 2014;82(3):695–708. https://doi.org/10.1016/j.neuron.2014.03.020 PMID: 24811386
13. Van Dijk KRA, Sabuncu MR, Buckner RL. The influence of head motion on intrinsic functional connectivity MRI. Neuroimage. 2012;59(1):431–8. https://doi.org/10.1016/j.neuroimage.2011.07.044 PMID: 21810475
14. Christoff K, Ream JM, Gabrieli JDE. Neural basis of spontaneous thought processes. Cortex. 2004;40(4–5):623–30. https://doi.org/10.1016/S0010-9452(08)70158-8
15. Gonzalez-Castillo J, Kam JWY, Hoy CW, Bandettini PA. How to Interpret Resting-State fMRI: Ask Your Participants. J Neurosci. 2021;41(6):1130–41. https://doi.org/10.1523/JNEUROSCI.1786-20.2020 PMID: 33568446
16. Noble S, Scheinost D, Constable RT. A guide to the measurement and interpretation of fMRI test-retest reliability. Curr Opin Behav Sci. 2021;40:27–32. https://doi.org/10.1016/j.cobeha.2020.12.012 PMID: 33585666
17. Noble S, Scheinost D, Constable RT. A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. Neuroimage. 2019;203:116157. https://doi.org/10.1016/j.neuroimage.2019.116157 PMID: 31494250
18. Hasson U, Nir Y, Levy I, Fuhrmann G, Malach R. Intersubject Synchronization of Cortical Activity During Natural Vision. Science. 2004;303(5664):1634–40. https://doi.org/10.1126/science.1089506
19. Wang J, Ren Y, Hu X, Nguyen VT, Guo L, Han J. Test-retest reliability of functional connectivity networks during naturalistic fMRI paradigms: Test-Retest Reliability of Naturalistic fMRI. Hum Brain Mapp. 2017;38(4):2226–41. https://doi.org/10.1002/hbm.23517
20. Eickhoff SB, Milham M, Vanderwal T. Towards clinical applications of movie fMRI. Neuroimage. 2020;217:116860. https://doi.org/10.1016/j.neuroimage.2020.116860 PMID: 32376301
21. Guo CC, Hyett MP, Nguyen VT, Parker GB, Breakspear MJ. Distinct neurobiological signatures of brain connectivity in depression subtypes during natural viewing of emotionally salient films. Psychol Med. 2016;46(7):1535–45. https://doi.org/10.1017/S0033291716000179 PMID: 26888415
22. Guo CC, Nguyen VT, Hyett MP, Parker GB, Breakspear MJ. Out-of-sync: disrupted neural activity in emotional circuitry during film viewing in melancholic depression. Sci Rep. 2015;5:11605. https://doi.org/10.1038/srep11605 PMID: 26112251
23. Hyett MP, Parker GB, Guo CC, Zalesky A, Nguyen VT, Yuen T, et al. Scene unseen: Disrupted neuronal adaptation in melancholia during emotional film viewing. Neuroimage Clin. 2015;9:660–7. https://doi.org/10.1016/j.nicl.2015.10.011 PMID: 26740919
24. Yang Z, Wu J, Xu L, Deng Z, Tang Y, Gao J, et al. Individualized psychiatric imaging based on inter-subject neural synchronization in movie watching. Neuroimage. 2020;216:116227. https://doi.org/10.1016/j.neuroimage.2019.116227 PMID: 31568871
25. O'Connor D, Potler NV, Kovacs M, Xu T, Ai L, Pellman J. The Healthy Brain Network Serial Scanning Initiative: A Resource for Evaluating Inter-Individual Differences and Their Reliabilities Across Scan Conditions and Sessions. GigaScience. 2017;6(2). https://doi.org/10.1093/gigascience/giw011
26. Zhang X, Liu J, Yang Y, Zhao S, Guo L, Han J, et al. Test-retest reliability of dynamic functional connectivity in naturalistic paradigm functional magnetic resonance imaging. Hum Brain Mapp. 2022;43(4):1463–76. https://doi.org/10.1002/hbm.25736 PMID: 34870361
27. Finn ES, Bandettini PA. Movie-watching outperforms rest for functional connectivity-based prediction of behavior. 2020. https://doi.org/10.1101/2020.08.23.263723

28. Li X, Friedrich P, Patil KR, Eickhoff SB, Weis S. A topography-based predictive framework for naturalistic viewing fMRI. 2022. https://doi.org/10.1101/2022.05.26.493420
29. Vanderwal T, Eilbott J, Finn ES, Craddock RC, Turnbull A, Castellanos FX. Individual differences in functional connectivity during naturalistic viewing conditions. Neuroimage. 2017;157:521–30. https://doi.org/10.1016/j.neuroimage.2017.06.027 PMID: 28625875
30. Ki JJ, Kelly SP, Parra LC. Attention Strongly Modulates Reliability of Neural Responses to Naturalistic Narrative Stimuli. J Neurosci. 2016;36(10):3092–101. https://doi.org/10.1523/JNEUROSCI.2942-15.2016 PMID: 26961961
31. Hasson U, Furman O, Clark D, Dudai Y, Davachi L. Enhanced intersubject correlations during movie viewing correlate with successful episodic encoding. Neuron. 2008;57(3):452–62. https://doi.org/10.1016/j.neuron.2007.12.009
32. Hasson U, Malach R, Heeger DJ. Reliability of cortical activity during natural stimulation. Trends Cogn Sci. 2010;14(1):40–8. https://doi.org/10.1016/j.tics.2009.10.011 PMID: 20004608
33. Kröll J-P, Friedrich P, Li X, Patil KR, Mochalski L, Waite L, et al. Naturalistic viewing increases individual identifiability based on connectivity within functional brain networks. Neuroimage. 2023;273:120083. https://doi.org/10.1016/j.neuroimage.2023.120083 PMID: 37015270
34. Tian L, Ye M, Chen C, Cao X, Shen T. Consistency of functional connectivity across different movies. Neuroimage. 2021;233:117926. https://doi.org/10.1016/j.neuroimage.2021.117926 PMID: 33675997
35. Bassett DS, Bullmore E. Small-World Brain Networks. Neuroscientist. 2006;12(6):512–23. https://doi.org/10.1177/1073858406293182
36. Reijneveld JC, Ponten SC, Berendse HW, Stam CJ. The application of graph theoretical analysis to complex networks in the brain. Clin Neurophysiol. 2007;118(11):2317–31. https://doi.org/10.1016/j.clinph.2007.08.010 PMID: 17900977
37. Bullmore E, Sporns O. Complex brain networks: graph theoretical analysis of structural and functional systems. Nat Rev Neurosci. 2009;10(3):186–98. https://doi.org/10.1038/nrn2575 PMID: 19190637
38. Deuker L, Bullmore ET, Smith M, Christensen S, Nathan PJ, Rockstroh B, et al. Reproducibility of graph metrics of human brain functional networks. Neuroimage. 2009;47(4):1460–8. https://doi.org/10.1016/j.neuroimage.2009.05.035 PMID: 19463959
39. Petrella JR. Use of graph theory to evaluate brain networks: a clinical tool for a small world?. Radiology. 2011;259(2):317–20. https://doi.org/10.1148/radiol.11110380
40. Telesford QK, Morgan AR, Hayasaka S, Simpson SL, Barret W, Kraft RA. Reproducibility of Graph Metrics in fMRI Networks. Frontiers in Neuroinformatics. 2010;4. https://doi.org/10.3389/fninf.2010.00117
41. Halchenko Y, Meyer K, Poldrack B, Solanky D, Wagner A, Gors J. DataLad: distributed system for joint management of code, data, and their relationship. JOSS. 2021;6(63):3262. https://doi.org/10.21105/joss.03262
42. Waites AB, Stanislavsky A, Abbott DF, Jackson GD. Effect of prior cognitive state on resting state networks measured with functional connectivity. Hum Brain Mapp. 2005;24(1):59–68. https://doi.org/10.1002/hbm.20069 PMID: 15382248
43. Vanderwal T, Kelly C, Eilbott J, Mayes LC, Castellanos FX. Inscapes: A movie paradigm to improve compliance in functional magnetic resonance imaging. Neuroimage. 2015;122:222–32. https://doi.org/10.1016/j.neuroimage.2015.07.069 PMID: 26241683
44. Esteban O, Markiewicz CJ, Blair RW, Moodie CA, Isik AI, Erramuzpe A, et al. fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat Methods. 2019;16(1):111–6. https://doi.org/10.1038/s41592-018-0235-4 PMID: 30532080
45. Avants B, Tustison NJ, Song G. Advanced Normalization Tools: V1.0. The Insight Journal. 2009. https://doi.org/10.54294/uvnhin
46. Jenkinson M, Bannister P, Brady M, Smith S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage. 2002;17(2):825–41. https://doi.org/10.1016/s1053-8119(02)91132-8 PMID: 12377157
47. Greve DN, Fischl B. Accurate and robust brain image alignment using boundary-based registration. Neuroimage. 2009;48(1):63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060 PMID: 19573611
48. Mandal S, Raimondo F, Sasse L, Komeyer V, Patil K, Hamdan S. juaml/junifer: v0.0.3. 2023. https://zenodo.org/record/8176569
49. Amft M, Bzdok D, Laird AR, Fox PT, Schilbach L, Eickhoff SB. Definition and characterization of an extended social-affective default network. Brain Struct Funct. 2015;220(2):1031–49. https://doi.org/10.1007/s00429-013-0698-0 PMID: 24399179
50. Buhle JT, Silvers JA, Wager TD, Lopez R, Onyemekwu C, Kober H, et al. Cognitive reappraisal of emotion: a meta-analysis of human neuroimaging studies. Cereb Cortex. 2014;24(11):2981–90. https://doi.org/10.1093/cercor/bht154 PMID: 23765157
51. Liu X, Hairston J, Schrier M, Fan J. Common and distinct networks underlying reward valence and processing stages: a meta-analysis of functional neuroimaging studies. Neurosci Biobehav Rev. 2011;35(5):1219–36. https://doi.org/10.1016/j.neubiorev.2010.12.012 PMID: 21185861
52. Sabatinelli D, Fortune EE, Li Q, Siddiqui A, Krafft C, Oliver WT, et al. Emotional perception: meta-analyses of face and natural scene processing. Neuroimage. 2011;54(3):2524–33. https://doi.org/10.1016/j.neuroimage.2010.10.011 PMID: 20951215
53. Bzdok D, Schilbach L, Vogeley K, Schneider K, Laird AR, Langner R, et al. Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. Brain Struct Funct. 2012;217(4):783–96. https://doi.org/10.1007/s00429-012-0380-y PMID: 22270812
54. Caspers S, Zilles K, Laird AR, Eickhoff SB. ALE meta-analysis of action observation and imitation in the human brain. Neuroimage. 2010;50(3):1148–67. https://doi.org/10.1016/j.neuroimage.2009.12.112 PMID: 20056149
55. Camilleri JA, Müller VI, Fox P, Laird AR, Hoffstaedter F, Kalenscher T, et al. Definition and characterization of an extended multiple-demand network. Neuroimage. 2018;165:138–47. https://doi.org/10.1016/j.neuroimage.2017.10.020 PMID: 29030105

56. Cieslik EC, Mueller VI, Eickhoff CR, Langner R, Eickhoff SB. Three key regions for supervisory attentional control: evidence from neuroimaging meta-analyses. Neurosci Biobehav Rev. 2015;48:22–34. https://doi.org/10.1016/j.neubiorev.2014.11.003 PMID: 25446951

57. Langner R, Eickhoff SB. Sustaining attention to simple tasks: a meta-analytic review of the neural mechanisms of vigilant attention. Psychol Bull. 2013;139(4):870–900. https://doi.org/10.1037/a0030694 PMID: 23163491

58. Rottschy C, Langner R, Dogan I, Reetz K, Laird AR, Schulz JB, et al. Modelling neural correlates of working memory: a coordinate-based meta-analysis. Neuroimage. 2012;60(1):830–46. https://doi.org/10.1016/j.neuroimage.2011.11.050 PMID: 22178808

59. Binder JR, Desai RH, Graves WW, Conant LL. Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb Cortex. 2009;19(12):2767–96. https://doi.org/10.1093/cercor/bhp055 PMID: 19329570

60. Spreng RN, Mar RA, Kim ASN. The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. J Cogn Neurosci. 2009;21(3):489–510. https://doi.org/10.1162/jocn.2008.21029 PMID: 18510452

61. Witt ST, Laird AR, Meyerand ME. Functional neuroimaging correlates of finger-tapping task variations: an ALE meta-analysis. Neuroimage. 2008;42(1):343–56. https://doi.org/10.1016/j.neuroimage.2008.04.025 PMID: 18511305

62. Hagberg A, Schult D, Swart P. Exploring Network Structure, Dynamics, and Function using NetworkX. In: 2008. https://conference.scipy.org/proceedings/SciPy2008/paper_2/

63. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychological Methods. 1996;1(1):30–46. https://doi.org/10.1037/1082-989x.1.1.30

64. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):420–8. https://doi.org/10.1037//0033-2909.86.2.420 PMID: 18839484

65. Cicchetti DV, Sparrow SA. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am J Ment Defic. 1981;86(2):127–37. PMID: 7315877

66. Müller R, Büttner P. A critical discussion of intraclass correlation coefficients. Statistics in Medicine. 1994;13(23–24):2465–76. https://doi.org/10.1002/sim.4780132310

67. Kong J, Gollub RL, Webb JM, Kong J-T, Vangel MG, Kwong K. Test-retest study of fMRI signal change evoked by electroacupuncture stimulation. Neuroimage. 2007;34(3):1171–81. https://doi.org/10.1016/j.neuroimage.2006.10.019 PMID: 17157035

68. Zhang H, Zhang Y-J, Duan L, Ma S-Y, Lu C-M, Zhu C-Z. Is resting-state functional connectivity revealed by functional near-infrared spectroscopy test-retest reliable?. J Biomed Opt. 2011;16(6):067008. https://doi.org/10.1117/1.3591020 PMID: 21721829

69. Masuda N, Sakaki M, Ezaki T, Watanabe T. Clustering Coefficients for Correlation Networks. Front Neuroinform. 2018;12:7. https://doi.org/10.3389/fninf.2018.00007 PMID: 29599714

70. Aron AR, Gluck MA, Poldrack RA. Long-term test-retest reliability of functional MRI in a classification learning task. Neuroimage. 2006;29(3):1000–6. https://doi.org/10.1016/j.neuroimage.2005.08.010 PMID: 16139527

71. Cao H, Plichta MM, Schäfer A, Haddad L, Grimm O, Schneider M, et al. Test-retest reliability of fMRI-based graph theoretical properties during working memory, emotion processing, and resting state. Neuroimage. 2014;84:888–900. https://doi.org/10.1016/j.neuroimage.2013.09.013 PMID: 24055506

72. Rikandi E, Mäntylä T, Lindgren M, Kieseppä T, Suvisaari J, Raij TT. Functional network connectivity and topology during naturalistic stimulus is altered in first-episode psychosis. Schizophr Res. 2022;241:83–91. https://doi.org/10.1016/j.schres.2022.01.006 PMID: 35092893

73. Zhang G, Liu X. Investigation of functional brain network reconfiguration during exposure to naturalistic stimuli using graph-theoretical analysis. J Neural Eng. 2021;18(5):10.1088/1741-2552/ac20e7. https://doi.org/10.1088/1741-2552/ac20e7 PMID: 34433142

74. Finn ES, Scheinost D, Finn DM, Shen X, Papademetris X, Constable RT. Can brain state be manipulated to emphasize individual differences in functional connectivity?. NeuroImage. 2017;160:140–51.

75. Hlinka J, Děchtěrenko F, Rydlo J, Androvičová R, Vejmelka M, Jajcay L, et al. The intra-session reliability of functional connectivity during naturalistic viewing conditions. Psychophysiology. 2022;59(10):e14075. https://doi.org/10.1111/psyp.14075 PMID: 35460523

76. DuPre E, Hanke M, Poline J-B. Nature abhors a paywall: How open science can realize the potential of naturalistic stimuli. Neuroimage. 2020;216:116330. https://doi.org/10.1016/j.neuroimage.2019.116330 PMID: 31704292

77. Saarimäki H. Naturalistic stimuli in affective neuroimaging: A review. Front Hum Neurosci. 2021;15:675068. https://doi.org/10.3389/fnhum.2021.675068

78. Adams RB Jr, Rule NO, Franklin RG Jr, Wang E, Stevenson MT, Yoshikawa S, et al. Cross-cultural reading the mind in the eyes: an fMRI investigation. J Cogn Neurosci. 2010;22(1):97–108. https://doi.org/10.1162/jocn.2009.21187 PMID: 19199419

79. Goh JOS, Leshikar ED, Sutton BP, Tan JC, Sim SKY, Hebrank AC, et al. Culture differences in neural processing of faces and houses in the ventral visual cortex. Soc Cogn Affect Neurosci. 2010;5(2–3):227–35. https://doi.org/10.1093/scan/nsq060 PMID: 20558408

80. Harada T, Mano Y, Komeda H, Hechtman LA, Pornpattananangkul N, Parrish TB, et al. Cultural influences on neural systems of intergroup emotion perception: An fMRI study. Neuropsychologia. 2020;137:107254. https://doi.org/10.1016/j.neuropsychologia.2019.107254 PMID: 31726067

81. Sneddon I, McKeown G, McRorie M, Vukicevic T. Cross-cultural patterns in dynamic ratings of positive and negative natural emotional behaviour. PLoS One. 2011;6(2):e14679. https://doi.org/10.1371/journal.pone.0014679 PMID: 21364739

82. Gal S, Coldham Y, Tik N, Bernstein-Eliav M, Tavor I. Act natural: Functional connectivity from naturalistic stimuli fMRI outperforms resting-state in predicting brain activity. Neuroimage. 2022;258:119359. https://doi.org/10.1016/j.neuroimage.2022.119359 PMID: 35680054

83. Schaefer A, Nils F, Sanchez X, Philippot P. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. Cognition & Emotion. 2010;24(7):1153–72. https://doi.org/10.1080/02699930903274322

84. Alink A, Schwiedrzik CM, Kohler A, Singer W, Muckli L. Stimulus predictability reduces responses in primary visual cortex. J Neurosci. 2010;30(8):2960–6. https://doi.org/10.1523/JNEUROSCI.3730-10.2010 PMID: 20181593

85. Koster-Hale J, Saxe R. Theory of mind: a neural prediction problem. Neuron. 2013;79(5):836–48. https://doi.org/10.1016/j.neuron.2013.08.020 PMID: 24012000

86. Mason MF, Norton MI, Van Horn JD, Wegner DM, Grafton ST, Macrae CN. Wandering minds: the default network and stimulus-independent thought. Science. 2007;315(5810):393–5. https://doi.org/10.1126/science.1131295 PMID: 17234951

87. Golland Y, Bentin S, Gelbard H, Benjamini Y, Heller R, Nir Y, et al. Extrinsic and intrinsic systems in the posterior cortex of the human brain revealed during natural sensory stimulation. Cereb Cortex. 2007;17(4):766–77. https://doi.org/10.1093/cercor/bhk030 PMID: 16699080

88. Andellini M, Cannatà V, Gazzellini S, Bernardi B, Napolitano A. Test-retest reliability of graph metrics of resting state MRI functional brain networks: A review. J Neurosci Methods. 2015;253:183–92. https://doi.org/10.1016/j.jneumeth.2015.05.020 PMID: 26072249

89. Liang X, Wang J, Yan C, Shu N, Xu K, Gong G, et al. Effects of different correlation metrics and preprocessing factors on small-world brain functional networks: a resting-state functional MRI study. PLoS One. 2012;7(3):e32766. https://doi.org/10.1371/journal.pone.0032766 PMID: 22412922

90. Murphy K, Birn RM, Handwerker DA, Jones TB, Bandettini PA. The impact of global signal regression on resting state correlations: are anti-correlated networks introduced?. Neuroimage. 2009;44(3):893–905. https://doi.org/10.1016/j.neuroimage.2008.09.036 PMID: 18976716

91. Schwarz AJ, McGonigle J. Negative edges and soft thresholding in complex network analysis of resting state functional connectivity data. Neuroimage. 2011;55(3):1132–46. https://doi.org/10.1016/j.neuroimage.2010.12.047 PMID: 21194570