

MASTER THESIS META-ANALYTICAL CONTRASTS: EVALUATION OF ROBUSTNESS AND VALIDITY

Submitted by

Vincent Küppers, B.Sc.

Matriculation number 2775702

vincent.kueppers@gmail.com

on the 20th of January 2021

In the Master Course

Translational Neuroscience

Faculty of Medicine

Heinrich-Heine University, Düsseldorf, Germany

First advisor: Dr. Robert Langner

Second advisor: Dr. Edna C. Cieslik-Köchling

INM-7: Brain and Behaviour

Team Behavioural Neuroscience

Research Centre Jülich, Jülich, Germany

Acknowledgments

I would like to thank Prof. Dr. Simon Eickhoff & Dr. Robert Langner for the opportunity to join the INM-7 and to write my master thesis there. I learned a lot during my stay at the institute, probably more than in any previous class. I would like to thank the whole Behavioural Neuroscience group for a very pleasant time in a scientifically inspiring environment. I would especially like to thank Dr. Veronika Müller for her scientific guidance, valuable feedback and huge support throughout the duration of my thesis. I would like to thank my friends and family for their support in these weird and dark times.

Finally, I would like to thank all the authors who kindly provided their additional results, without which this analysis would not have been possible. Namely, Prof. Dr. Cristina Forn Frías, Dr. Salvatore Campanella, Dr. Marcel Daamen, Yu Fukuda, Prof. Dr. Ute Habel, Prof. Dr. Kathrin Koch, Dr. Ian Harding, Kyesam Jung, Dr. Jakob Kaminski, Prof Dr. Tricia Z. King, Prof. Dr. Axel Krug, Dr. Jacob Lahr and Dr. Lora Minkova, Xiao-you Zhang and Prof. Dr. Lin Li, Dr. Anna Miró Padilla, Prof. Dr. Frank Schneider, Prof. Dr. Florian Schlagenhauf, Prof. Dr. Marion Smits, Dr. Mara Rocca, Dr. Harm Jan van der Horn and Dr. Yuan Zhou.

Table of Contents

ACKNOWLEDGMENTS	I
AFFIRMATION IN LIEU OF OATH.....	IV
LIST OF ABBREVIATIONS.....	V
LIST OF FIGURES.....	VI
LIST OF TABLES	VI
LIST OF SUPPLEMENTARY FIGURES AND TABLES	VI
ABSTRACT	1
INTRODUCTION	2
META-ANALYSIS	2
CONTRAST BETWEEN TWO META-ANALYSES.....	3
META-ANALYTICAL COMPARISONS: TWO DIFFERENT APPROACHES	4
VALIDITY, ROBUSTNESS AND SAMPLE SIZES	6
AIM AND OBJECTIVES	8
MATERIALS AND METHODS.....	9
TASK OF INTEREST	9
LITERATURE-BASED META-ANALYSIS.....	10
<i>Literature search</i>	10
<i>Inclusion and exclusion criteria</i>	10
<i>Coding of coordinates</i>	11
ACTIVATION LIKELIHOOD ESTIMATION (ALE).....	13
<i>ALE contrast analyses</i>	14
<i>Evaluation of meta-analytical contrasts</i>	14
<i>Evaluation of contrast baseline/ control conditions</i>	15
COMPARISON OF THE META-ANALYTICAL RESULTS TO THE N-BACK NETWORK DERIVED FROM A LARGE INDIVIDUAL FMRI STUDY	16
<i>HCP</i>	16
<i>Subject and group level GLM modelling</i>	17
LARGE-SAMPLE-SIMULATED META-ANALYSIS	18
<i>Drawing and computing of studies</i>	18
<i>Coordinate extraction</i>	19
<i>Meta-analyses</i>	19
MEASURE OF SIMILARITY AND VALIDITY	19
<i>Jaccard similarity coefficient</i>	20
<i>Sensitivity and Specificity</i>	21
<i>Voxel-wise comparison</i>	21

<i>ROI-wise comparison</i>	22
ANATOMICAL LABELLING	22
CODE AND DATA AVAILABILITY	23
RESULTS	24
EVALUATION OF META-ANALYTICAL CONTRASTS	24
EVALUATION OF CONTRAST BASELINE/ CONTROL CONDITIONS	28
EVALUATION OF POWER IN META-ANALYTICAL CONTRASTS	31
<i>Comparison of Cmeta with MCexp (within iteration)</i>	31
<i>Comparison of Cmeta with literature MCexp</i>	33
DISCUSSION	34
REFERENCE NETWORKS: META-ANALYSIS ACROSS CONTRASTS AND LARGE-SAMPLE N-BACK CONTRAST ..	34
META-ANALYTICAL CONTRAST EVALUATION	35
INFLUENCE OF TYPE OF CONTRASTING CONDITION	37
ARE EXPERIMENTS CONTRASTING VS. REST SUITED FOR A CBMA?	39
MORE STUDIES ALONE ARE NOT ENOUGH FOR A ROBUST META-ANALYTICAL CONTRAST	41
HOW MANY EXPERIMENTS ARE RECOMMENDABLE FOR A CMETA?	41
WHICH EFFECTS CAN BE SHOWN IN A CMETA?	42
GENERAL DISCUSSION	43
<i>Potential problems on the experimental level</i>	44
<i>Potential problems on the meta-analytical level</i>	45
<i>Masking and multiple comparison correction in meta-analytical contrast</i>	45
LIMITATIONS	46
OUTLOOK	47
CONCLUSION	48
BIBLIOGRAPHY	50
SUPPLEMENTARY DATA	58
OVERVIEW OF LITERATURE DATASET	58
ADDITIONAL LITERATURE-BASED META-ANALYSES RESULTS	66
SUPPLEMENTARY BIBLIOGRAPHY	68

Affirmation in lieu of oath

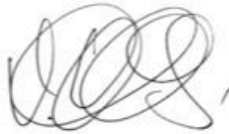
Eidesstattliche Versicherung / Affirmation in lieu of oath

Ich versichere, dass ich die vorliegende Masterarbeit selbständig verfasst und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt, nur die angegebenen Quellen benutzt habe. Ich habe die Stellen gekennzeichnet, die ich wörtlich oder inhaltlich den benutzten Quellen entnommen habe. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen.

I hereby declare that I have authored the present Master thesis independently and without making use of any means other than those indicated, only using the sources indicated. I marked the passages which I have taken literally or with regards to content from the sources used. The work has not been submitted in the same or similar form to any other examination authority.

I hereby confirm that the printed version of this thesis is identical to the digital version handed in to the examination authority.

Köln, 20/01/2021



(Ort, Datum, Unterschrift des Kandidaten/der Kandidatin /
Place, Date, Signature Candidate)

List of Abbreviations

ALE	Activation likelihood estimation
ANOVA	Analysis of variance
Bsl.	Baseline condition
Cmeta	Contrast between two meta-analyses
COI	Contrast of interest
EEG	Electroencephalography
fMRI	Functional magnetic resonance imaging
HCP	Human connectome project
L	Left brain hemisphere
MEG	Magnetoencephalography
MCexp	Meta-analysis across contrasts on experimental level
MDD	Major depressive disorder
R	Right brain hemisphere
ROI	Region of interest
SCZ	Schizophrenia
Sign.	Significant
TPR	True positive rate, sensitivity
TNR	True negative rate, specificity
WM	Working memory

Anatomic abbreviations

ANG	Angular gyrus
FL	Frontal pole
IFG	Inferior frontal gyrus
INS	Insular cortex
IPL	Inferior parietal lobule
IPS	Inferior parietal sulcus
LOC	Lateral occipital cortex
MFG	Middle frontal gyrus
OFC	Orbitofrontal gyrus

PcG	Paracingulate gyrus
PCUN	Precuneus cortex
PreCG	Precentral gyrus
SFG	Superior frontal gyrus
SMA	Juxtapositional lobule / supplementary motor area
SMG	Supramarginal gyrus
SPL	Superior parietal lobule

List of Figures

FIGURE 1. LETTER N-BACK PARADIGM. THE RED CIRCLE INDICATES THE STIMULUS TO BE IDENTIFIED.....	9
FIGURE 2. FLOWCHART OF LITERATURE-BASED META-ANALYSES.....	12
FIGURE 3. META-ANALYSIS ACROSS 2-BACK > 0-BACK EXPERIMENTS (62).....	24
FIGURE 4. LARGE-SAMPLE 2-BACK > 0-BACK NETWORK.	26
FIGURE 5. BRAIN REGIONS REVEALED BY CONTRAST BETWEEN META-ANALYSES (2-BACK > REST/ BASELINE) > (0-BACK > REST/ BASELINE) (CMETA) AND BY META-ANALYSES ACROSS 2-BACK > 0-BACK EXPERIMENTS (MCEXP)	27
FIGURE 6. BRAIN REGIONS REVEALED BY CONTRAST BETWEEN META-ANALYSES (2-BACK > 0-BACK) > (1-BACK > 0-BACK) (CMETA) AND BY META-ANALYSES ACROSS 2-BACK > 1-BACK EXPERIMENTS (MCEXP).....	29
FIGURE 7. META-ANALYTICAL CONTRAST SIMULATION USING A LARGE SAMPLE.....	32

List of Tables

TABLE 1. BRAIN REGIONS INVOLVED IN 2-BACK VERSUS 0-BACK NETWORKS REVEALED BY MCEXP AND CMETA	24
TABLE 2. SIMILARITY OF BRAIN NETWORKS.....	28
TABLE 3. BRAIN REGIONS INVOLVED IN 2-BACK VERSUS 1-BACK NETWORKS REVEALED BY MCEXP AND CMETAS.....	30

List of Supplementary Figures and Tables

SUPPLEMENTARY TABLE 1. OVERVIEW OF EXPERIMENTS INCLUDED IN THE DIFFERENT META-ANALYSES.	58
SUPPLEMENTARY FIGURE 1. A) META-ANALYSIS ACROSS 0-BACK > REST/ BASELINE EXPERIMENTS (21); B) META-ANALYSIS ACROSS 2-BACK > REST/ BASELINE EXPERIMENTS (31).....	66
SUPPLEMENTARY FIGURE 2. BRAIN REGIONS REVEALED BY CONTRAST BETWEEN META-ANALYSES (2-BACK > REST/ BASELINE) > (0-BACK > REST/ BASELINE) (CMETA) AND BY LARGE SAMPLE 2-BACK > 0-BACK CONTRAST	67

SUPPLEMENTARY FIGURE 3. REDUCED 2-BACK > 0-BACK: CONTRAST BETWEEN TWO META-ANALYSES

(CMETA) VS. META-ANALYSIS ACROSS CONTRASTS (MCEXP)	67
--	----

Abstract

Background: Meta-analytical contrasts have become a widely used tool in recent years, adding powerful capabilities to the Activation Likelihood Estimation (ALE) meta-analysis algorithm (Eickhoff et al., 2011; Laird, Fox, et al., 2005). Through integration of existing neuroimaging literature, it allows to show commonalities and differences between different subject groups (e.g. Yapple et al., 2019), task modalities (e.g. Langner & Eickhoff, 2013; Rottschy et al., 2012) or facets of larger cognitive constructs (e.g. Langner et al., 2018; Morelli et al., 2015). Importantly, it allows to test hypotheses that have not been or can't be addressed at single study level. However, little is known about the validity of meta-analytical contrasts, lacking systematic empirical investigations.

Methods: Focusing on a specific task contrast (n-back), a contrast between two meta-analyses ((2-back > baseline) > (0-back > baseline)) was compared with a meta-analysis across contrasts on experimental level (2-back > 0-back). This was done in a literature-based approach (i.e. a traditional meta-analysis) and using a large sample of subject level contrasts to simulate differently powered meta-analytical contrasts.

Results & Conclusions: Contrasts between two meta-analyses show regional significant stronger convergence to a lesser degree compared to a meta-analysis across contrasts. Regions that are identified can be interpreted with relative certainty similar to regions found in a meta-analysis across contrasts on experimental level, i.e. likelihood of false positives is low. However, it is not recommended to interpret the absence of regions because the sensitivity is relatively low. Experiments with contrasts against a control condition are found to be more suited than experiments with contrasts against rest/passive baseline conditions for the computation of a contrast between two meta-analyses. Finally, an increase in the number of experiments used is associated with better sensitivity and similarity, but it seems that above a number of 26 experiments this increase is reduced with a higher likelihood of getting false positives.

Introduction

Human brain mapping is the effort to discover the relationships between the anatomy and function of the human brain. Over the past 30 years, neuroimaging has emerged as a promising method in the study of brain organization, with thousands of publications each year. Functional magnetic resonance imaging (fMRI) is a noninvasive imaging technique that allows researchers to test cognitive paradigms while simultaneously scanning a participant's brain. This so-called task-based fMRI has been and continues to be used to study the neural substrate of various neurological and psychiatric disorders, cognitive and socio-affective processes, and more.

Although fMRI has enabled numerous discoveries and contributed to a better understanding of the brain, it also has its drawbacks. The temporal resolution is relatively low compared to other techniques (e.g. MEG, EEG), the cost of scanning subjects is high, and the results are probably highly dependent on a number of factors.

Those factors include sample size, which is a hotly debated issue (Button et al., 2013; Cremers et al., 2017; Geuter et al., 2018; Poldrack et al., 2017; Turner et al., 2018) and in-scanner time of individual participants, which has been debated in recent years (Nee, 2019). Another factor is the different (pre-)processing and analysis of fMRI data. Different laboratories use different software and different pipelines (Botvinik-Nezer et al., 2020; Carp, 2012). Not to mention, that different scanners may produce different results (Jovicich et al., 2006). Furthermore, when using task-based fMRI, experimental flexibility is an important issue. Thus, different cognitive task paradigms with different variations can be used to study the same cognitive process. Variations in these factors can lead to differences in results and sometimes make replication nearly impossible (Bossier et al., 2020).

Meta-analysis

A way to overcome the issue of spurious findings, the problem of low power, analytical and experimental flexibility in individual fMRI studies are meta-analyses (Cremers et al., 2017). A meta-analysis on neuroimaging data can show which results, which brain regions, which cognitive networks show convergence in brain activations across experiments. A widely used and successful method to perform neuroimaging meta-analyses is the Activation Likelihood Estimation (ALE) (Eickhoff et al., 2009, 2012; Turkeltaub et al., 2002) algorithm.

Numerous neuroimaging ALE meta-analyses have been performed to identify the neural substrate of various psychological constructs, e.g. working memory (Rottschy et al., 2012), vigilant attention (Langner & Eickhoff, 2013), social cognition (Bzdok et al., 2012) and language processing (Ferstl et al., 2008). ALE meta-analysis are also conducted more specifically across a single task paradigm, e.g. n-back (Owen et al., 2005), Stroop task (Laird, McMillan, et al., 2005), Go/No-go tasks (Simmonds et al., 2008). Other applications include the identification of neural differences in different subject groups, e.g. in age (Heckner et al., 2020; Yapple et al., 2019) or patient groups, e.g. schizophrenia (Minzenberg et al., 2009).

Meta-analyses have the power not only to confirm and reveal robust brain networks, but also to show where there is no convergence across results, potentially indicating replicability issues. Negative results are an important topic in science today and fortunately receiving more attention. In a meta-analysis, Müller and colleagues (2017) found that the current neuroimaging literature does not reveal robust abnormalities in major depressive disorder (MDD) patients compared to healthy controls. Another example comes from Chuan-Peng and colleagues (2020), who used an ALE meta-analysis to demonstrate the lack of a neural basis for a universal beauty center in the brain.

Despite these qualities of neuroimaging meta-analyses, they are also affected by potential pitfalls. A meta-analysis can only be performed on the basis of published literature. Accordingly, selective reporting of results (publications bias), p-hacking, etc. are possible sources of errors (Müller et al., 2018). The quality of the experiments on which a meta-analysis is based is therefore of crucial importance.

Contrast between two meta-analyses

In addition to the classical meta-analysis that looks for convergent results across experiments, there is also the possibility of calculating contrasts and conjunctions between meta-analyses similar to individual fMRI experiments.

An ALE contrast analysis, hereafter referred to as contrast between two meta-analyses (Cmeta), was first introduced by Laird et al. (2005). It was developed to statistically compare two ALE maps by testing the null hypothesis that both sets of foci are uniformly distributed. Often results of meta-analyses are compared and conclusions are drawn on the basis that a region appears in one meta-analysis but not in another. However, such a difference may simply be due to the fact that in one meta-analysis the region is just below

the significance threshold, while in the other it is just above the threshold. Therefore, a test was introduced to formally test whether differences are significant. The tool has been further developed and is currently based on a label-exchange permutation test (Eickhoff et al., 2011).

Since its introduction, the contrast between two meta-analyses has been used to examine a wide variety of brain-behavior relationships. For example, it can be used to compare different subtypes of larger psychological concepts. One example is the comparison of verbal and non-verbal WM experiments, which confirmed the role of the left brain area 44/45 in speech functions (Rottschy et al., 2012). In addition to comparing different task modalities, such as auditory versus visual paradigms (Langner & Eickhoff, 2013), different tasks can also be compared that are supposed to isolate the same cognitive process, e.g. in WM: n-back vs. Sternberg tasks (Rottschy et al., 2012). A comparison of meta-analyses also provides the opportunity to compare different groups of subjects. A study by Yapple et al. (2019) showed age-related differences in the convergence of brain regions associated with the n-back task. They found a decline in convergence of prefrontal cortex engagement with age. Another example is the study of the human self-regulation system by Langner and colleagues (2018). They were able to uncover differences in cognitive emotion regulation and cognitive action regulation through a meta-analytical comparison and confirm these differences through additional analyses.

Thus, contrasts between two meta-analyses can play a valuable role in testing current neurocognitive theories, establishing new theories, and uncovering the neural substrate of these.

Meta-analytical comparisons: two different approaches

In general, differences between 2 groups or conditions (e.g. *condition A > condition B*) can be investigated with meta-analyses in 2 different ways. Either one does a meta-analysis across experiments, that look at the contrast of interest (COI) on an experimental level, i.e. one includes in the analysis those experiments that do a subtraction analysis on e.g. *condition A* vs. *condition B*. This type of analysis is called meta-analysis across contrasts on experimental level (MCexp).

A contrast on an experimental level can be obtained by a subtraction analysis of different brain scans. A typical experiment would include a cognitive task with the cognitive component of interest (e.g. *condition A*) and a control task with the same cognitive

processes involved but without the cognitive component of interest (e.g. *condition B*). In the subtraction, the brain activity during the cognitive task is compared to activity during the control task. Instead of the control task, one can also use the implicit baseline, i.e. resting-state measurement (or a fixation baseline). This depends on what the investigator is looking for (Price et al., 1997).

In a MCexp the contrast of interest is calculated always on an experimental level and the meta-analysis just shows which brain regions show convergence in brain activations for this contrast. In other words, the results of a meta-analysis can be interpreted as the convergence of brain activations of a specific COI.

However, there is also another way to investigate a COI in a meta-analysis. As opposed to the MCexp (e.g. *condition A > condition B*) a contrast can also be calculated on the meta-analytical level. This is done by performing two independent meta-analyses for both conditions of interest (e.g. *condition A > baseline* and *condition B > baseline*). And in a second step the contrast between two meta-analyses (Cmeta) is calculated. The resulting contrast is again a contrast between two conditions (e.g. *(condition A > baseline) > (condition B > baseline)*) but derived in a totally different way. Importantly, a Cmeta is a conceptual different contrast, as in the later approach only the differences in convergence of brain activations are evaluated and not the convergence of differences of brain activations (Müller et al., 2018). In other words, a MCexp shows the convergence of brain activations between conditions across different experiments. A Cmeta on the other hand, shows stronger convergence of one meta-analysis in comparison to another meta-analysis.

Performing a MCexp across a specific contrast, be it a task or group contrast, is probably always the preferred choice. This is because a Cmeta probably means a higher loss of information compared to a MCexp, due to the conceptual differences explained above. However, this strictly limits meta-analyses to contrasts studied multiple times in the literature. Cmeta may be the answer to overcome this limitation and ask questions that have not been asked at the experimental level or cannot be asked. For example, if one wants to compare two clinical groups, like patients with schizophrenia (SCZ) and patients with MDD. Contrasts between SCZ vs. controls and contrasts between MDD vs. controls probably abound in the literature. However, there may be too few studies that report a direct contrast between SCZ vs. MDD. This would mean that it would not be possible to compute a MCexp, but it would be possible to compute a Cmeta. While there are many

interesting questions that can be addressed with Cmeta, little is known about its validity and robustness.

Validity, robustness and sample sizes

While the results of an ALE meta-analysis of neuroimaging data are fairly straightforward to interpret, it is a bit difficult to interpret a contrast between two meta-analyses. By definition, it is a statistical comparison of two ALE maps. This comparison shows in which brain regions convergence is significantly stronger in one meta-analysis compared to another meta-analysis. The question that arises here, and which has not really been empirically investigated yet, is what this stronger convergence states. To what extent does this conceptually different contrast reflect the same thing shown by a MCexp?

Furthermore, the extent to which a meta-analytical contrast identifies regions that are really relevant for the process is unknown. Is there a possibility that regions in a Cmeta show significantly stronger convergence, but these are not at all related to supposedly studied process?

A major factor that shapes meta-analyses are the experiments used to conduct each analysis. Experiment selection is an essential part of meta-analyses and it is important not to put apples and oranges together unless one is studying approximately round fruits. A strength of meta-analyses is the analysis of heterogeneous data and finding commonalities. Potentially, a problem could arise if the dataset is based solely on one type of contrast and the reported results do not reflect the totality of the process it purports to represent. This could be the case if a contrast is depicting a multitude of processes, as it is the case in contrasts between complex conditions versus a rest baseline (Price et al., 1997; Stark & Squire, 2001). The baseline condition is crucial here, since it means that more processes are subtracted out (higher level baseline) of the contrast or remain in it (lower level baseline). Reported results of a contrast against a very low baseline may not reflect the totality of all processes. This could invalidate the meta-analysis calculated from it. This is true for meta-analyses in general but is equally relevant when one performs a contrast between meta-analyses. Therefore, it is essential to be sure that the experiments are appropriate for an ALE meta-analysis and that the reported results contain what they purport to.

In the design of every experiment, the sample size is an essential factor for replicability and inference. In the field of fMRI research low statistical power, due to small number

sample sizes, is subject to an ongoing discourse (Button et al., 2013; Cremers et al., 2017; Desmond & Glover, 2002; Poldrack et al., 2017). As for individual fMRI studies, sample size is also important for meta-analyses of neuroimaging studies. It is assumed that the number of experiments required for a meta-analysis is between 17 and 20. In a simulation study, Eickhoff and colleagues (2016) showed that this is the minimum number of experiments needed to detect robust medium sized effects and to ensure that the results are not driven by single experiments. However, the influence of sample size and what is optimal for Cmetas is not yet known. It goes without saying that there should be at least 17 experiments in each meta-analysis between which the contrast is calculated. However, we actually do not know how many experiments are optimal for the current contrast analysis. Are 17 experiments sufficient to obtain valid results, or does it need to be more? Therefore, an empirical assessment of sample size in Cmetas is necessary to provide a basis to help authors of meta-analyses to make an informed decision.

Aim and Objectives

A foundation for understanding the neural basis for cognitive processes is provided through decades of research by the neuroimaging community. Contrast analyses at the individual study level are one of the widely used practices. However, study-level contrasts are limited by numerous factors. At a meta-analytical level, a contrast can potentially overcome some of the limitations and address novel and interesting questions. However, little is known about the validity and robustness of meta-analytical contrasts.

The aim of this project is to evaluate and to uncover the factors that influence meta-analytical contrasts. In the first part of this investigation a meta-analysis across contrasts on the experimental level (MCexp) will be evaluated against the contrast between two meta-analyses (Cmeta).

Next, a second set of literature-based meta-analyses will be conducted to test the relevance that the type of condition against which the contrast is conducted has on meta-analytical contrasts.

Then, a large-sample-simulated meta-analysis will be used to create different scenarios to test the influence of the number of experiments on the validity of meta-analytical contrasts.

This is an exploratory study to gain insight into how to interpret meta-analytical contrasts, which meta-analytical approaches yield robust results, and ultimately help in the design of future meta-analytical studies.

Materials and Methods

Task of interest

The question of validity and robustness of meta-analytical contrasts was investigated using the example of a specific cognitive task. The “n-back” paradigm (Kirchner, 1958) was chosen because it is widely used in the literature, well established and performed in a relatively similar way (Owen et al., 2005). This working memory paradigm consists of a series of stimuli presented to a participant. The participant is asked to identify the stimuli that were presented to her/ him n trials back (Figure 1). Typically, n is equal to 1, 2 or 3. In the case of a control task (e.g. 0) the participant is simply asked to identify a stimulus that is defined prior to the trial. Other control conditions could be either a passive fixation baseline, or resting state measurements (classified as rest) or a more engaging baseline, e.g. flashing fixation cross (classified as baseline). To investigate the research question, contrasts between the following conditions were selected: 2-back > 0-back, 2-back > rest/ baseline, 0-back > rest/ baseline, 2-back > 1-back, 2-back > rest/ baseline, 2-back > 0-back, 1-back > rest/ baseline, 1-back > 0-back.

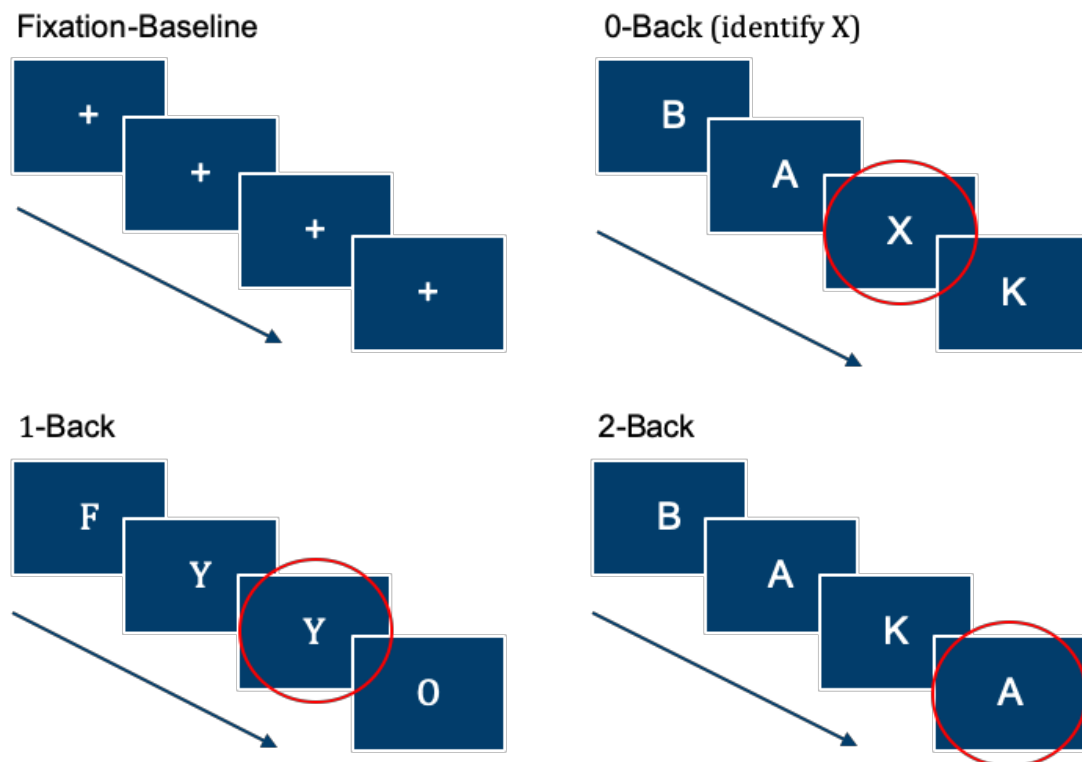


Figure 1. Letter n-back paradigm. The red circle indicates the stimulus to be identified.

Literature-based meta-analysis

Literature search

The data for the literature-based meta-analyses was collected in three different ways. The initial dataset was based on the n-back studies collected by Rottschy et al. (2012). Secondly, reference tracing was carried out from other previous meta-analytical studies by Langner & Eickhoff (2013); Yaple et al. (2019) and Mencarelli et al. (2019). In addition to this method, two online search engines, "pubmed" (<https://pubmed.ncbi.nlm.nih.gov/>) and "web of science" (<http://webofknowledge.com/>) were searched for the following keywords: ("fMRI" OR "functional MRI" OR "functional magnetic resonance imaging") AND ("n-back" OR "0-back" OR "zero-back"). Additional filters were chosen to include only human studies and a publication date between 2012/01/01 - 2020/01/01. After reading the abstracts, a total of 254 articles were further analyzed. A second effort was made by contacting authors directly via E-Mail with a request to contribute additional results, if available. This was done when none of the desired coordinates were reported in the publication, but the study design met the criteria. Or when some contrasts were reported (e.g., studies that were already part of the previous dataset) and the analysis/study design implied that other contrasts were analyzed but not reported in the publication. The e-mail address was determined by a "Google" search of the first/corresponding author, if no valid e-mail was found, the last author was contacted instead. This may be the case, for example, if the first author has moved to another research group or a position in industry. If the first request was not answered, a friendly reminder was sent a few weeks later.

Inclusion and exclusion criteria

In accordance with the general guidelines (Müller et al., 2018), the experiments were considered eligible if they contained coordinates of whole-brain contrasts (not ROI), used standard analysis procedures, the subjects were healthy (healthy control groups and healthy variations of the general population; i.e. no diagnosed diseases) and were over 18 years of age. An experiment was included if any of the contrasts of interest (COI) was reported, all other contrasts were excluded (e.g. 3-back > 0-back). Furthermore, only positive contrasts (i.e. 2-back > 0-back) were considered, since only a small fraction reported deactivations (i.e. 0-back > 2-back). Experiments that reported different variations of the "n-back" paradigm (e.g. spatial or verbal tasks; different types of stimuli such as visual or auditory) were included. Experiments were excluded when an

intervention was part of the experimental design (e.g. drug testing or working memory training). However, if a baseline measurement (without any intervention) was attempted, the resulting contrast would meet the requirements.

Coding of coordinates

If coordinates were available in the peer-reviewed studies, the coordinates of the contrasts of interest were coded. As most studies reported only the main effect of working memory (i.e. 2-back > 0-back), and not always the contrasts against rest/ baseline, potential authors were kindly asked to contribute their additional data, if available. The additional, not peer-reviewed results were provided in different formats. Some were output tables of the used analysis software (SPM/ FSL), with the local maxima of the clusters, others were (unthresholded) activation t-maps (in Nifti-format). To extract the relevant foci, the methods described in the corresponding studies were used (i.e. correction method, cluster extent). If no or unclear information about the coordinate extraction was available, all of the provided coordinates (or local peak maxima) were extracted and coded. If the results were sent as activation maps, the first 10 peak coordinates per cluster and a cluster extent threshold of $k = 10$ were used. For the coordinate extraction the SPM Anatomy Toolbox v2.2 (Eickhoff et al., 2005) was used.

The control conditions were classified as “rest” if the control conditions for the computed contrast were either resting state measurements or a passive fixation baseline (i.e. a stationary fixation cross). The category “baseline” was chosen if the participants were asked to pay attention to a flashing fixation cross, or passively perceived the stimuli (e.g. the trials in the same frequency).

A reported contrast (i.e. a set of coordinates corresponding to a contrast between two conditions) is counted as one experiment. Thereby one study (i.e. one publication) can report more than one experiment (for example one study reporting 2-back > 0-back, 2-back > rest and 0-back > rest in the same subject group). If more than one experiment per group was reported for different conditions all experiments were selected as they were included in different meta-analyses. However, if for one subject group the same contrast was reported twice, e.g. with a different n-back paradigm but between the same conditions (e.g. 2-back > 0-back visual n-back and 2-back > 0-back auditory n-back) only one of them was included in the dataset. This was done to reduce effects driven by a specific group of subjects.

In total the dataset of all experiments (including received results) for both approaches (contrast comparison, investigation of baseline condition) consisted of 108 studies and 170 experiments. More precisely, the dataset consisted of 62 experiments with the contrast 2-back > 0-back, 20 experiments with the contrast 2-back > 1-back, 31 experiments with the contrast 2-back > rest/ baseline, 21 experiments with the contrast 0-back > rest/ baseline, 19 experiments with the contrast 1-back > 0-back and 17

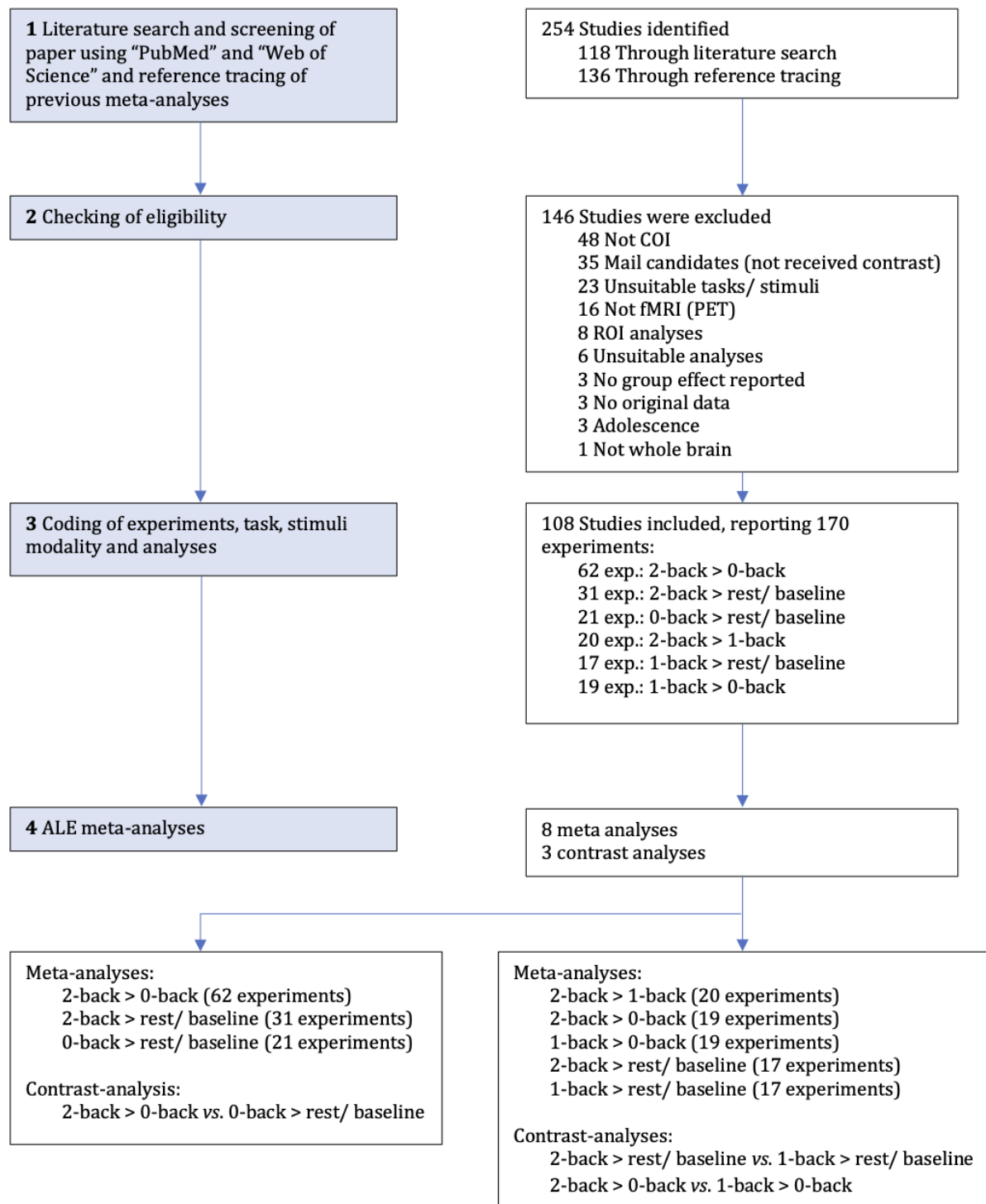


Figure 2. Flowchart of literature-based meta-analyses.

experiments with the contrast 1-back > rest/ baseline. For 10 groups all the 3 COI for the contrast comparison (2-back > 0-back; 2-back > rest/ baseline, 0-back > rest/ baseline) was available. Detailed information about the included experiments can be found in the Supplementary Information (Supplementary Table 1) and about the analysis steps Figure 2.

Activation Likelihood Estimation (ALE)

The coordinate-based meta-analyses were calculated with the ALE algorithm (Eickhoff et al., 2009, 2012; Turkeltaub et al., 2002, 2012). For the computation the Python-based NiMARE implementation of the ALE algorithm was used (Salo et al., 2020).

In ALE, the reported foci from each study are treated as the centers of 3D Gaussian probability distributions reflecting spatial uncertainty. The width of the distributions depends on the number of subjects from a study. Larger sample sizes may reflect more reliable spatial accuracy and therefore result in narrower widths. For each experiment, all probability distributions of all reported foci are combined into a modeled activation map (MA). This results in one MA map per experiment with associated probabilities at each voxel. When the Gaussian probability distributions of different foci overlap, the probabilities associated with the voxel within that overlap are given only by the foci closest to that voxel. This is done to avoid the summation of within-group effects (Turkeltaub et al., 2012). Then, for each voxel, the union of all probabilities across MA maps (i.e. individual experiments) is calculated. The ALE scores derived in this way show convergence across experiments rather than individual foci (Eickhoff et al., 2009).

To distinguish true convergence from random convergence, the ALE maps are tested against a null distribution. This null distribution is analytically derived, based on a nonlinear histogram algorithm and reflects the random spatial associations (Eickhoff et al., 2012). In this algorithm, each MA-map is converted into a histogram. The histogram shows the probability of the occurrences of all possible MA-values (i.e. activation probability) by joining all voxels with the same MA-value (including zeros) in single histogram-bins. Under the assumption of spatial independence, the normalized histograms are combined by cycling through all the non-zero bins of two histograms. The combined histogram is then joined with a third histogram and so on until a final histogram representing a probabilistic distribution of ALE-values is derived.

To correct for multiple comparisons, an empirical null distribution is generated using a permutation approach. A cluster-wise family-wise error corrected (cFWE) threshold at $p < 0.05$ with a cluster-forming threshold of $p < 0.001$ is used. At each of the 10000 permutations the foci are exchanged with randomly selected voxels within a grey matter mask, ALE values are calculated and the maximum cluster size per permutation is recorded. A cluster is treated significant if its size is larger than or equal to the 95 % (cluster extent threshold of $p < 0.05$) percentile of the generated clusters. Reported peaks reflect the local maxima z-scores of the uncorrected map at voxel-level.

ALE contrast analyses

As introduced by Laird et al. (2005) and modified by Eickhoff et al. (2011), the statistical comparison between two ALE Maps was performed using the NiMARE *ALESubtraction* algorithm implemented in version 0.0.3. (Salo et al., 2020). The algorithm was used to compute the contrast analysis, one-sided (activations only) across all voxels. Specifically, the two original ALE scores are calculated for two groups and then a difference score is calculated for each voxel (i.e. voxel-wise subtraction). In a permutation-based approach, all experiments (contributing to one of the contrasts) are pooled and randomly divided in two groups of experiments with the same size as the original groups. The ALE scores of these pooled groups are calculated and the differences in the ALE score for each voxel are recorded. By repeating this label exchange process 10000 times, an empirical null distribution of differences in ALE scores is formed against which the observed differences in ALE scores are tested. The voxels are then tested for significant differences at a threshold of $P > .95$ (i.e. the observed probability is equal or higher than 95 % chance level). In addition, the contrast map is inclusively masked with the significant cFWE-corrected main effect of the respective group. Note that masking is not performed within the original NiMARE implementation but conducted subsequently in order to draw inference from the contrast analysis. On a second note, the technique described does not correct for multiple comparisons.

Evaluation of meta-analytical contrasts

First, it will be investigated to what extent a contrast between two meta-analyses reflects what is shown in a meta-analysis across contrasts on experimental level (MCexp). For the MCexp experiments are used that contrast the two conditions of interest (i.e. contrast between *condition A > condition B*). To calculate the contrast between two meta-analyses (Cmeta) requires a two-step procedure: first two meta-analyses are calculated, one for

each condition of interest against a resting baseline (meta-analysis 1: *condition A > implicit baseline* and meta-analysis 2: *condition B > implicit baseline*). Second, a contrast between these two meta-analyses is performed. Thus, the same contrast is calculated as in the first meta-analysis (meta-analysis across contrasts), but in a conceptually different way.

The current project includes all experiments reporting the 2-back > 0-back contrast for the MCexp (Figure 2). Conceptually, this meta-analysis reveals all those regions that consistently show stronger activation in 2-back vs. 0-back across experiments. Additionally, two meta-analyses with all experiments with the 0-back > rest/ baseline (*condition B > implicit baseline*) contrasts and a second with the contrasts 2-back > rest/ baseline (*condition A > implicit baseline*) are performed and then contrasted with each other on the meta-analytical level (i.e. (2-back > rest/ baseline) > (0-back > rest/ baseline)) (Figure 2). This contrast shows conceptually all those regions that show more convergence in one meta-analysis than in the other. This conceptually different contrasts (MCexp and Cmeta) are then compared and similarities and differences evaluated. This is done in a visual descriptive approach and by calculating different similarity metrics (which will be described in more detail later).

Evaluation of contrast baseline/ control conditions

A second approach was chosen to investigate the influence that the type of condition against which the contrast is performed has on the results of Cmeta. In contrast to the first approach, where meta-analyses were calculated that included experiments against a rest/ fixation baseline (*condition A/B > implicit baseline*), Cmeta between two meta-analyses across experiments contrasting against a control condition (*condition A/B > condition C*) was additionally performed. The MCexp is computed across all 2-back > 1-back experiments (*condition A > condition B*) and one Cmeta between the 2-back > rest/ baseline (*condition A > implicit baseline*) and 1-back > rest/ baseline experiments (*condition B > implicit baseline*) and a second Cmeta between 2-back > 0-back (*condition A > condition C*) and 1-back > 0-back experiments (*condition B > condition C*) were computed. Thereby three contrasts are computed, all in a conceptual different way (Figure 2).

The collected sample (Supplementary Table 1) was reduced because the number of experiments in each meta-analysis varied widely. The reduced sample included all 20 “2-back > 1-back” experiments, all 19 “1-back > 0-back” experiments, 19 hand-matched “2-

back > 0-back” experiments, all 17 “1-back > rest/ baseline” experiments and 17 hand-matched “2-back > rest/ baseline” experiments. The matching of studies was performed in order to get pairs of relatively similar studies. First all experiments were included of studies where both conditions were reported (i.e. 2-back > 0-back and 1-back > 0-back; or 2-back > rest/ bsl. and 1-back > bsl.). The remaining experiments were matched to the respective lower level contrasts primarily by sample size and modality, secondly by stimulus and age group (Supplementary Table 1). Note that the subsampling in the meta-analysis across 2-back > rest/ baseline and meta-analysis across 2-back > 0-back was conducted just once.

Comparison of the meta-analytical results to the n-back network derived from a large individual fMRI study

In addition to the above described meta-analysis across 2-back > 0-back experiments a second reference contrast was computed using a large sample from the HCP dataset (Van Essen et al., 2013). This reference contrast is not based on a meta-analysis, but is a single, individual experiment looking at the contrast of interest in a large sample of subjects. This allows the meta-analytical contrasts to be compared to an independent contrast.

HCP

The HCP 1200 Subjects Release (S1200) contains data of 1206 subjects, 657 subjects are female with a mean age of 30.01 (Std. deviation: 3.522) and 549 subjects are male with a mean age of 28.48 (Std. deviation: 3.665). The HCP dataset includes a wide range of related subjects, twins and relatives. This means a lot of the subjects come from the same family. Within the whole release there are 457 unique Family IDs. The release contains 8 different task paradigms. For the comparison to the meta-analytical results the working memory (WM) n-back task fMRI data was used. After exclusion of subjects with known quality issues (HCP Data Release Updates: Known Issues and Planned Fixes - Connectome Data Public - HCP Wiki; HCP Subjects with Identified Quality Control Issues (QC_Issue Measure Codes Explained)) and including only subjects who completed the n-back task and scored > 50 % Accuracy in 2-back condition the whole sample consisted of 1020 (435 unrelated) subjects.

The Data was pre-processed according to the minimal pre-processing pipeline and analyzed with the FSL FEAT module (Glasser et al., 2013).

Subject and group level GLM modelling

As current literature, especially older studies are based entirely on volume based processing the minimally preprocessed volume based time-series (Glasser et al., 2013) for the WM task were used to create subject-level WM maps of the COIs (2BK, 0BK, 2BK-0BK) using a modified version of the HCPpipeline scripts (Barch et al., 2013; <https://github.com/Washington-University/HCPpipelines>). Following a temporal filtering, using a high pass filter of 200 the time series were smoothed using a FWHM of 8mm. This kernel width was chosen as it is found to be the most frequently used in the experiments constituting the literature dataset. The GLM model fitting was done according to the Volume-Based Analysis of the HCP pipeline, eight predictors were included for each stimulus type in each n-back condition (Stimuli: faces, places, body, tools; conditions: 0-back, 2-back). The predictors ranged from the presentation of the cue to the final trial of a stimulus block (27.5 s). These blocked predictors were convolved with a double “canonical” hemodynamic response function (HRF). The temporal derivatives of each predictor were included as regressors of no interest to compensate for slice timing variability and HRF delay across regions (Barch et al., 2013). 12 movement regressors were included (6 motion parameter estimates from the rigid-body transformation and their temporal derivatives) as further confounds of no interest. Three linear contrasts were computed on basis of this GLM model: 2-back vs. 0-back, 2-back vs. fixation, 0-back vs. fixation. For subject-level effect estimates a fixed-effects analysis was conducted across both runs, within subjects. The HCP pipeline scripts are based on the FSL FEAT module (Woolrich et al., 2001, 2004).

The group-level GLM was calculated using a Python NiPype (Esteban et al., 2020) based workflow using the FSL interface (Woolrich et al., 2004). The analysis scripts are based on an example task based FSL workflow created by Esteban et al., 2019. The group level effects were estimated using the FSL FLAME (FMRIB's Local Analysis of Mixed Effects) module conducting a one-sample t-test between a random draw of unrelated subjects (435). As the estimated effects for the HCP reference networks were very large after voxel-wise FWE correction, resulting in a single cluster per map, effect size maps were computed. The computation of the effect-size (Cohen's d) estimates was done by dividing the group-level contrast of parameter estimate (COPE) by the square root of the estimated variance of parameter estimate (VARCOPE) divided by the square root of the sample size (N) 435 (see Equation 1) (Poldrack et al., 2017).

Equation 1. Effect size estimation.

$$Cohen's\ d\ map = \frac{\frac{COPE}{\sqrt{VARCOPE}}}{\sqrt{N}}$$

The resulting effect size map was thresholded at $d = 0.5$ (medium effect size) to get only medium to high effects according to Cohen (Cohen, 2013). This is done because ALE meta-analyses detect usually detect only moderate to strong effects (Eickhoff et al., 2016; Salimi-Khorshidi et al., 2009).

Large-sample-simulated meta-analysis

Analogous to the meta-analytical contrast evaluation using a literature derived dataset a second investigation was conducted to evaluate the influence of sample size on a meta-analytical contrast and validate the from literature derived findings. Therefore, a large sample (HCP) was used to simulate (mimicking a realistic meta-analysis scenario) differently powered meta-analyses. This was done by creating a pool of subject-level contrasts, as described above and drawing from this sample different studies. Group (study) level contrasts were computed and subsequently one MCexp and one Cmeta was computed. This was repeated with different numbers of studies (K) and 100 iterations for each variation. The Cmeta was compared to the MCexp within the iteration and to the literature meta-analysis across 2-back > 0-back network.

Drawing and computing of studies

The subject level analysis is equal to the above described pipeline for the large sample contrast. From the above described n-back sample, for 3 subjects the data was unavailable resulting in 1017 subject level contrasts. For each of the 100 iterations, from the pool of subjects, K 'studies' were drawn. A study is analogous to the literature in that it is a set of group-level contrasts based on a particular set of subjects. The number of studies for a meta-analysis was varied between 17 and 38 in steps of 3. The maximum of K was determined by the absolute sample size. The minimum was chosen according to the minimum recommendations for an ALE meta-analysis by Eickhoff and colleagues (2016). The number of subjects per study (N) are based on the 2-back > 0-back literature dataset (i.e. the 62 experiments with the contrast 2-back > 0-back). This means a study could contain between 8 and 84 subjects. The individual subsamples were drawn from a uniform distribution (between 8 and 84), but such that the overall mean was equal to the mean number of subjects of the literature dataset (26.24 ± 0.25). The subject drawing was

done without replacement and preventing 2 subjects in one study with the same family ID (i.e. every study contains only unrelated subjects). This was done to prevent heritability related effects within a study (Blokland et al., 2011). For every study the group level effects for all 3 contrasts (2-back vs. 0-back, 2-back vs. fixation, 0-back vs. fixation) were estimated as described for the large sample ($n = 435$) contrast.

Coordinate extraction

Three different thresholding methods were used for the group-level contrasts. Cluster-level FWE $p < 0.05$ corrected (0.001 cluster-forming threshold), uncorrected $p < 0.001$ and voxel-level FWE corrected 0.05. These were chosen as they represent the three most frequent methods found in the literature dataset. For every study one thresholding method was randomly determined. If no significant activations remained after thresholding, another method was chosen. In a few cases (in 6 of 22000 studies) of small subject sizes (e.g. $N = 8$) activations were not seen in the 2-back > 0-back contrast. In these cases, only the significant contrasts were picked. Coordinates were extracted from the randomly chosen thresholded maps. All local maxima were extracted with a minimum distance of 16 mm. These criteria were chosen to cover the whole range of large clusters and limiting the overall number of peaks by choosing a relatively wide peak distance.

Meta-analyses

Based on the extracted coordinates, 3 meta-analyses were computed for every iteration within every variation of K. As in the literature-based approach, 1 meta-analysis across all 2-back > 0-back contrasts were computed and a contrast between the meta-analysis across 2-back vs. fixation and across 0-back vs. fixation. The ALE meta-analyses and contrast between two-metanalyses computations were performed as described above.

Measure of similarity and validity

Comparisons are performed between the Cmetas and reference networks. The different reference networks are regarded as the validation networks. Similarity (as Jaccard similarity coefficient), sensitivity and specificity as described below are regarded measures of criterion validity.

The literature-based dataset was used for two meta-analytical investigations. In the first analysis, the similarity between a Cmeta and two reference networks (MCexp and large-sample contrast) was evaluated. In the second analysis, two Cmetas were compared to one MCexp reference network to examine the influence of contrasting conditions on

Cmeta. Quantitative similarity, sensitivity and specificity between the Cmeta and the reference networks was assessed by a voxel-wise comparison across the entire brain. In addition, a ROI-wise comparison was performed to provide a more qualitative comparison.

In the large-sample-simulated meta-analyses, the Cmetas were compared with two reference networks. The first comparison with the MCexp within iteration, derived from the same studies. The second comparison with the literature-based MCexp across 2-back > 0-back contrasts. As before, voxel-wise and ROI-wise similarity, sensitivity and specificity were calculated. This was performed for each iteration. Subsequently, the measured similarities were averaged over the iterations. Lastly, multiple one-way ANOVAs with subsequent post-hoc tests were conducted to assess the effect of sample size (K) on measured similarity, sensitivity and specificity.

Jaccard coefficient, sensitivity and specificity were computed on the Python based nilearn package (0.6.2) (Abraham et al., 2014). The statistical significance tests were conducted with the stats package in the R programming language (4.0.3).

Jaccard similarity coefficient

The percent overlap of two fMRI networks was computed using the Jaccard similarity coefficient (Jaccard, 1901). This measurement is similar to the similarity coefficient of Dice and Sørensen (Dice, 1945; Sørensen, 1948) and was introduced as a measure for reproducibility in fMRI studies by calculating the percent overlap of activations (Maitra, 2010).

The Jaccard coefficient as a comparison of two ALE/ activation maps (Cmeta vs. reference) is the intersection divided by the union of all significant voxels/ regions of both maps. Following the binary classification test between the two maps, the intersection can also be described as all true positives (TP) (sign. voxels/ regions in Cmeta and in reference). The union can be described as the sum of TP, false negatives (FN) (sign. voxels/ regions not in Cmeta but in reference) and false positives (FP) (sign. voxels/ regions in Cmeta but not in reference) (see Equation 2).

Equation 2. Jaccard similarity coefficient

$$Jaccard = \frac{TP}{TP + FN + FP}$$

A Jaccard coefficient of 1 would indicate a perfect overlap of both maps. All sign. and not sign. voxels/ regions would be identical in that case. The minimal Jaccard similarity coefficient is 0.

Sensitivity and Specificity

Sensitivity and specificity are widely used measures in medicine (Altman & Bland, 1994). While in medical diagnostics those measures assess the validity of a test, in this study they are an assessment of the validity of the Cmeta.

Sensitivity measures the proportion of sign. voxels/ regions (positives) that are correctly identified in the Cmeta. Correctly in this context means as in the reference map. Based on the measures of binary classification, as described above, the sensitivity (true positive rate, TPR) can be computed as true positives divided by the sum of true positives and false negatives (see Equation 3).

Equation 3. Sensitivity, true positive rate.

$$TPR = \frac{TP}{TP + FN}$$

The sensitivity range is from 0 to 1. A TPR of 1 would indicate that the Cmeta is depicting all sign. regions/ voxels present in the reference.

Specificity measures the proportion of not sign. voxels/ regions (negatives) that are correctly identified in the Cmeta. The specificity (true negative rate, TNR) can be computed as all true negatives (TN) (not sign. voxels/ regions in both maps) divided by the sum of true negatives and false positives (see Equation 4).

Equation 4. Specificity, true negative rate

$$TNR = \frac{TN}{TN + FP}$$

A TNR of 1 would mean that all not sign. voxels/ regions in the reference are also not sign. in the Cmeta. The closer the TNR is to 0, the more regions are falsely identified.

Voxel-wise comparison

The first measure of similarity was done by comparing the whole brain. First, both maps (Cmeta and references) are binarized, such that all sign. voxels are coded as 1 and all non-sign. voxels as a 0. Then, both maps are compared voxel by voxel. From this, the binary classifications (TP, FP, TN, FN) can be calculated as described above. Based on the

classifications the Jaccard-coefficient (Equation 2), sensitivity (Equation 3) and specificity (Equation 4) were calculated.

The resulting coefficients indicate the extent to which the Cmeta map matches the reference maps. This voxel-wise similarity is based on the premise that the maps to be compared should ideally be identical. A Jaccard similarity coefficient of 1 would indicate the same voxels are sign. in both maps.

ROI-wise comparison

However, for the evaluation of the meta-analytical contrast it is not necessary that the voxels of both maps are identical. It is much more important that all regions covered in the reference contrasts are also represented in the Cmeta. Thus, it is more a question of an overlap of the regions than a similarity of the whole maps. Therefore, an additional ROI-wise approach was chosen.

To identify ROIs, a combined atlas of the Brainnetome 246 and Diedrichsen 28 region probabilistic atlas of the human cerebellum was used (J. Diedrichsen et al., 2011; Jörn Diedrichsen et al., 2009; Fan et al., 2016). This combined 1.25 mm parcellation (BNA274) (<http://www.brainnetome.org/resource/>) was resampled using FSL FLIRT (FMRIB's Linear Image Registration Tool) to 2 mm isomorphic (Greve & Fischl, 2009; M. Jenkinson & Smith, 2001; Mark Jenkinson et al., 2002). Using this combined parcellations the qualitative difference between the Cmetas and the reference networks was calculated by comparing if within a given ROI significant voxels were detected.

Given the measures of similarities, as described above, the Cmeta maps were compared with the reference maps. If at least 1 voxel is significant in each of the two images (reference and Cmeta) in an ROI, the ROI is counted as a true positive (TP), if a ROI contains significant voxel only in the reference but not in the Cmeta it is considered a false negative (FN), if a ROI contains in both networks no significant voxels it is considered a true negative (TN) and if only in the Cmeta sign. voxels are within the ROI it is considered a false positive (FP). From this, the Jaccard similarity coefficient (Equation 2), sensitivity (Equation 3) and specificity (Equation 4) were calculated.

Anatomical labelling

All labelling of the anatomic regions was done by referring to the up to now histologically defined brain regions as reported within the SPM Anatomy Toolbox Version 3 (Eickhoff et al., 2005). Cytoarchitectonic locations were reported if the probabilities exceeded 5 %.

Detailed information on the cytoarchitectonic maps can be found in the respective publications on Area 44 and Area 45 (K. Amunts et al., 1999; Katrin Amunts et al., 2004), on Area hIP1 (IPS), Area hIP2 (IPS), Area hIP3 (IPS), Area 7PC (SPL), Area 7P (SPL) and Area 7A (SPL) (Scheperjans, Eickhoff, et al., 2008; Scheperjans, Hermann, et al., 2008), on Area hIP6 (IPS) and Area hIP8 (IPS) (Richter et al., 2019) and on Area 6mr / preSMA (Ruan et al., 2018). In addition, information about the microanatomical region was given for all resulting brain regions.

Code and data availability

All scripts and data used in the analyses will be gladly provided on request.

Results

Evaluation of meta-analytical contrasts

A meta-analysis across all 2-back > 0-back experiments was performed to calculate a literature derived reference network for a meta-analytical contrast. In this meta-analysis across contrasts on experimental level (MCexp), significant convergence was found in the left Broca's Region (44/ 45) and smaller cluster in the right Area 45, left and right middle frontal gyrus (MFG), left and right IPS/ SPL, paracingulate gyrus and (pre-)SMA, bilateral cerebellum, left frontal pole (FL), left caudate and thalamus and a bilateral cluster spanning over the posterior part of the orbitofrontal cortex (OFC) and the anterior insular cortex (aINS) (Table 1, Figure 3)

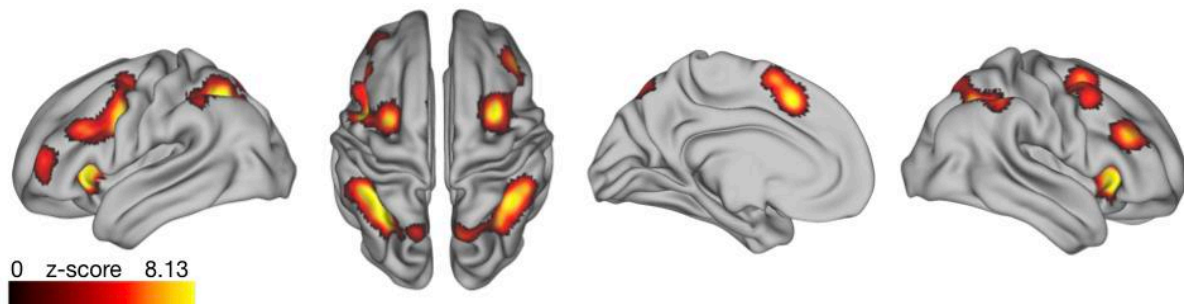


Figure 3. Meta-analysis across 2-back > 0-back experiments (62).

Table 1. Brain regions involved in 2-back versus 0-back networks revealed by MCexp and Cmeta

Contrast	Cluster Voxel	Local Peaks (Macroanatomical location)	Cytoarchitectonic Location - Probabilities at local maximum (%)	MNI coordinates			z-score
				x	y	z	
MCexp: 2-back > 0-back	1622	L PreCG	Area 44, 25.2 %	-42	4	28	8.13
		L MFG		-28	0	54	7.75
		L IFG (p. Opercularis)	Area 44, 8 %	-46	18	26	6.38
			Area 45, 15.1 %				
	1556	L SMG, posterior division	Area hIP1 (IPS), 56.8 %	-34	-48	38	8.13
			Area hIP2 (IPS), 14.4 %				
			Area hIP3 (IPS), 6.4 %				
	1469	R ANG	Area hIP1 (IPS), 76.1 %	42	-48	42	8.13
			Area hIP2 (IPS), 22%				
		R LOC, superior division	Area hIP1 (IPS), 19.3 %	32	-58	46	7.01
			Area hIP3 (IPS), 22.6 %				
			Area hIP6 (IPS), 6.9 %				
	1023	R PCUN	Area 7A (SPL), 11.9 %	10	-66	52	6.16
			Area 7P (SPL), 5.5 %				
			Area 6mr / preSMA, 9,7 %	4	14	46	8.13

		L SMA	Area 6mr / preSMA, 72.4 %	-4	8	58	6.75
789		R MFG	-	30	4	54	8.13
		R MFG	-	40	10	46	5.64
575		R MFG	Area 45, 20.8 %	46	32	26	6.32
546		R OFC	-	36	22	-8	8.13
532		L Cerebellum (Crus I)	-	-32	-60	-34	8.08
520		R Cerebellum (Crus I)	-	32	-62	-30	8.13
420		L INS	-	-30	20	-6	8.13
336		L FL	-	-36	48	8	5.55
201		R Cerebellum Lobule VI	-	10	-76	-24	8.08
177		L LOC superior division	Area 7P (SPL), 37.7 % Area 7A (SPL), 22.5 %	-12	-70	56	5.38
173		L Cerebellum Lobule VI	-	-8	-76	-28	7.51
126		L Caudate Nucleus	-	-16	-2	16	5.04
		L Thalamus	-	-12	-8	8	4.57
Cmeta: (2-back > rest/bsl.) vs. (0-back > rest/bsl.)	425	L SMG posterior division	Area hIP2 (IPS), 32.2 % Area hIP3 (IPS), 31.4 % Area hIP1 (IPS), 27.1 % Area 7PC (SPL), 9.3 %	-40	-46	42	8.21
		L SMG, posterior division	Area hIP1 (IPS), 47.4 % Area hIP2 (IPS), 25.7 % Area hIP3 (IPS), 20.1 % Area 7PC (SPL), 6.9 %	-38	-50	42	3.72
		L LOC, superior division	Area hIP6 (IPS), 35.5 % Area hIP3 (IPS), 14.0 % Area hIP8 (IPS), 5.9 %	-28	-68	44	2.69
	125	R ANG	Area hIP1 (IPS), 69.6% Area hIP2 (IPS), 30.1 %	40	-48	40	2.89
	95	R OFC	-	32	24	-8	2.71
	63	L OFC	-	-34	24	-6	2.27
	51	R SFG	-	28	4	62	2.24
	3	L MFG	-	-50	14	36	2.16
	1	LOC, superior division	Area 7A (SPL), 63.9 % Area hIP3 (IPS), 11.3 %	26	-62	58	1.7

Note: MCexp = Meta-analysis across contrasts on experimental level, Cmeta = Contrast between two meta-analyses; L = left hemisphere, R = right hemisphere; Cytoarchitectonic areas: Area 44 and Area 45 (K. Amunts et al., 1999; Katrin Amunts et al., 2004); Area hIP1 (IPS) and Area hIP2 (IPS) (Choi et al., 2006); Area hIP3 (IPS), Area 7PC (SPL), Area 7P (SPL) and Area 7A (SPL) (Scheperjans, Eickhoff, et al., 2008; Scheperjans, Hermann, et al., 2008); Area hIP6 (IPS) and Area hIP8 (IPS) (Richter et al., 2019); Area 6mr / preSMA (Ruan et al., 2018).

A second reference network of the contrast of interest was computed in only one individual experiment but from a large sample of subjects ($n = 435$). The contrast of 2-back > 0-back shows higher activations in a large bilateral frontal-parietal network. The frontal part of the network spans bilaterally from the orbitofrontal cortex, across the MFG (including DLPFC), parts of the SFG (including Broca's region (44/ 45)) and the insular cortex. On the medial surface including the paracingulate gyrus and the medial part of the SFG and anterior part of the supplementary motor cortex. Subcortical regions include the bilateral striatum (including ventral/dorsal caudate, nucleus accumbens, putamen) and parts of the thalamus. Parietal regions include the bilateral superior and inferior parietal lobule. Additionally, on the right hemisphere, is the right middle and inferior temporal gyrus, posterior division showing activation differences. Large parts of the cerebellum (not illustrated in the figure) show also activation differences. Lastly, some activation differences included regions in the brain stem (Supplementary Figure 2 A),B); Figure 4)

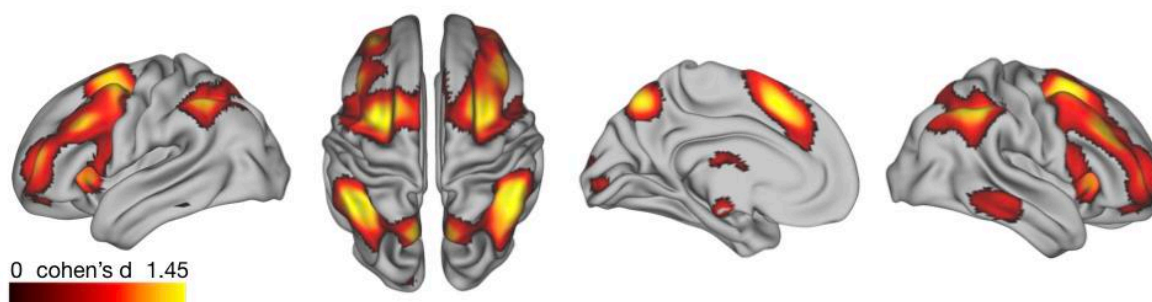


Figure 4. Large-sample 2-back > 0-back network.

The same contrast (2-back > 0-back) was computed on a meta-analytical level by taking a conceptual different approach. First, two meta-analyses were conducted across experiments, one meta-analysis across 2-back > rest/ baseline (31 experiments) contrasts and one meta-analysis across 0-back > rest/ baseline (21 experiments) contrasts. Significant convergence in these meta-analyses can be found in the Supplementary Figure 1.

Secondly, the ALE map of the experiments across 2-back > rest/ baseline was contrasted with the ALE map of the experiments across 0-back > rest/ baseline. Clusters of significant stronger convergence are found in the left and right IPS/ SPL, left and right aINS/ posterior OFC, right SFG and a very small cluster in the left MFG (Table 1, Figure 5).

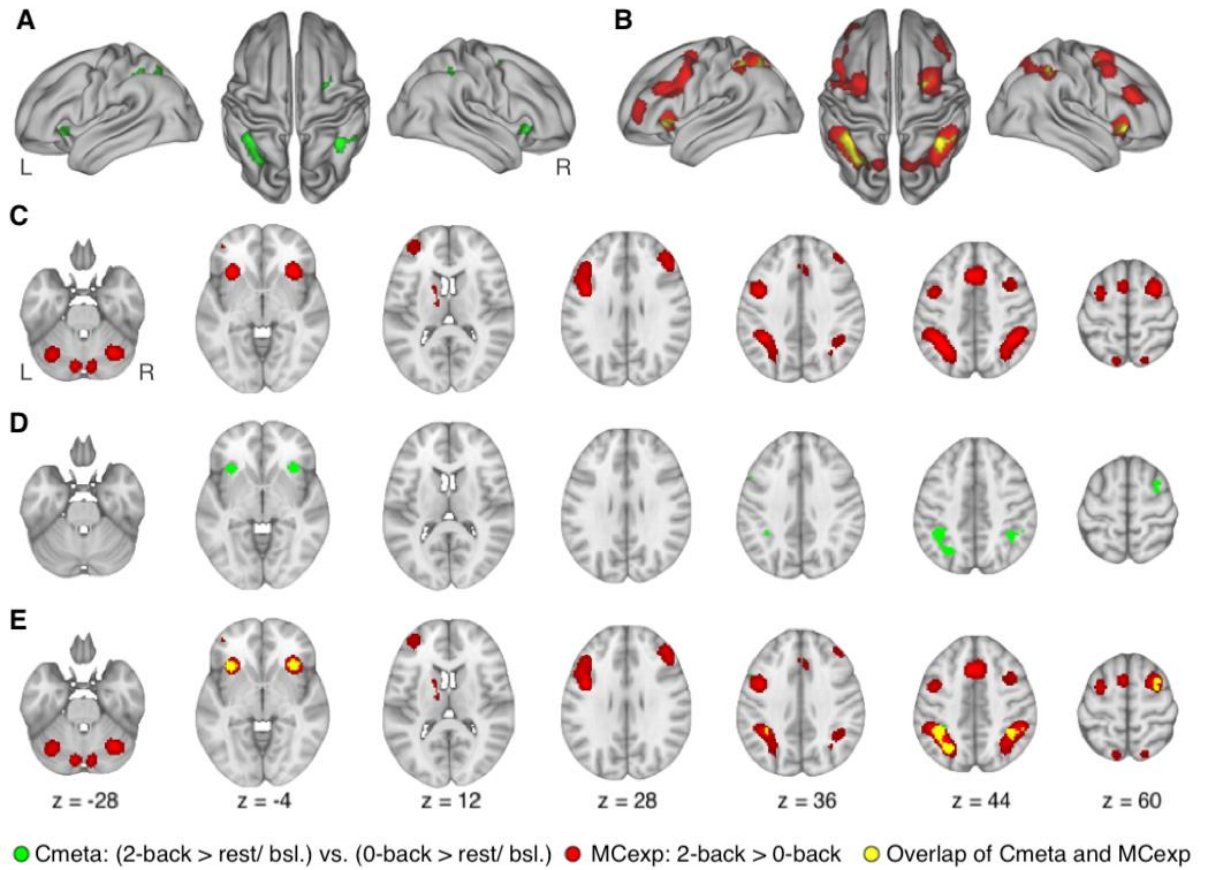


Figure 5. Brain regions revealed by contrast between meta-analyses (2-back > rest/ baseline) > (0-back > rest/ baseline) (Cmeta) and by meta-analyses across 2-back > 0-back experiments (MCexp). (A) and (D) show Cmeta. (B) and (E) show MCexp in red; Cmeta in green; Overlap of MCexp and Cmeta in yellow. (C) MCexp. Top row, cortex maps. Axial slices in MNI space.

Assessing the similarity of the contrast between two meta-analyses (Cmeta) (whole sample) with the two reference networks, no regions showing significant stronger convergence in the Cmeta lay outside the MCexp and the large-sample contrast. The Cmeta network is generally smaller, and several frontal regions that are found in the MCexp (and large-sample contrast) are not found in the Cmeta. The common regions are the left and right IPL (rostradorsal, caudal)/ SPL (intraparietal, lateral), left and right aINS (or posterior OFC), right SFG (dorsolateral area 6) and the left MFG (inferior frontal junction) (See Figure 5 for overlap to MCexp and Supplementary Figure 2 for overlap to large-sample contrast).

The calculated percent-overlap between the MCexp and the Cmeta is greater than the percent-overlap between the large-sample contrast and the Cmeta. This is true for the whole brain calculated Jaccard coefficient (0.074 vs. 0.019) and the qualitative, BNA 274 based comparison (Jaccard coefficient 0.265 vs. 0.155) (Table 2). In all comparisons the

sensitivity is equal to the Jaccard index, since there are no false positives. Thus, the specificity is perfect (TNR = 1).

Table 2. Similarity of brain networks

Reference-network	Contrast-network	Whole brain			BNA 274		
		Jaccard	TPR	TNR	Jaccard	TPR	TNR
MCexp: 2-back > 0-back	Cmeta: (2-back > rest/bsl.) vs. (0-back > rest/bsl.)	0.074	0.074	1	0.265	0.265	1
Large sample (HCP): 2-back > 0-back	Cmeta: (2-back > rest/bsl.) vs. (0-back > rest/bsl.)	0.019	0.019	1	0.155	0.155	1
MCexp: 2-back > 1-back	Cmeta: (2-back > 0-back) vs. (1- back > 0-back)	0.051	0.06	0.999	0.441	0.517	0.98
MCexp: 2-back > 1-back	Cmeta: (2-back > rest/bsl.) vs. (1-back > rest/bsl.)	0.001	0.001	1	0.035	0.035	1

Note: MCexp = Meta-analysis across contrasts on experimental level, Cmeta = Contrast between two meta-analyses; TPR = True positive rate (sensitivity), TNR = True negative rate (specificity)

Evaluation of contrast baseline/ control conditions

To investigate the effect of the baseline condition in a meta-analytical contrast 5 meta-analyses were conducted. A meta-analysis across 2-back > 1-back experiments was conducted (MCexp). As in the former approach this contrast should serve as a reference contrast for the contrasts between meta-analyses (Cmeta). In this meta-analysis significant convergence was found in the left Broca's area, left SPL, in the left and right IPS, at the paracingulate gyrus and in the left and right cerebellum (see Table 3, Figure 6).

Two different Cmetas were computed, one using experiment 0-back as the control condition and one using rest/baseline as the control condition. The first Cmeta was computed between the meta-analyses across 2-back > 0-back (19 experiments, 401 subjects) and across 1-back > 0-back (19 experiments, 404 subjects) and the second Cmeta between the meta-analyses across 2-back > rest/baseline (17 experiments, 388 subjects) and across 1-back > rest/baseline (17 experiments, 389 subjects).

The contrast analysis comparing the 2-back > rest/ baseline with the 0-back > rest/ baseline experiments showed significant stronger convergence in one very small cluster in the left cerebellum.

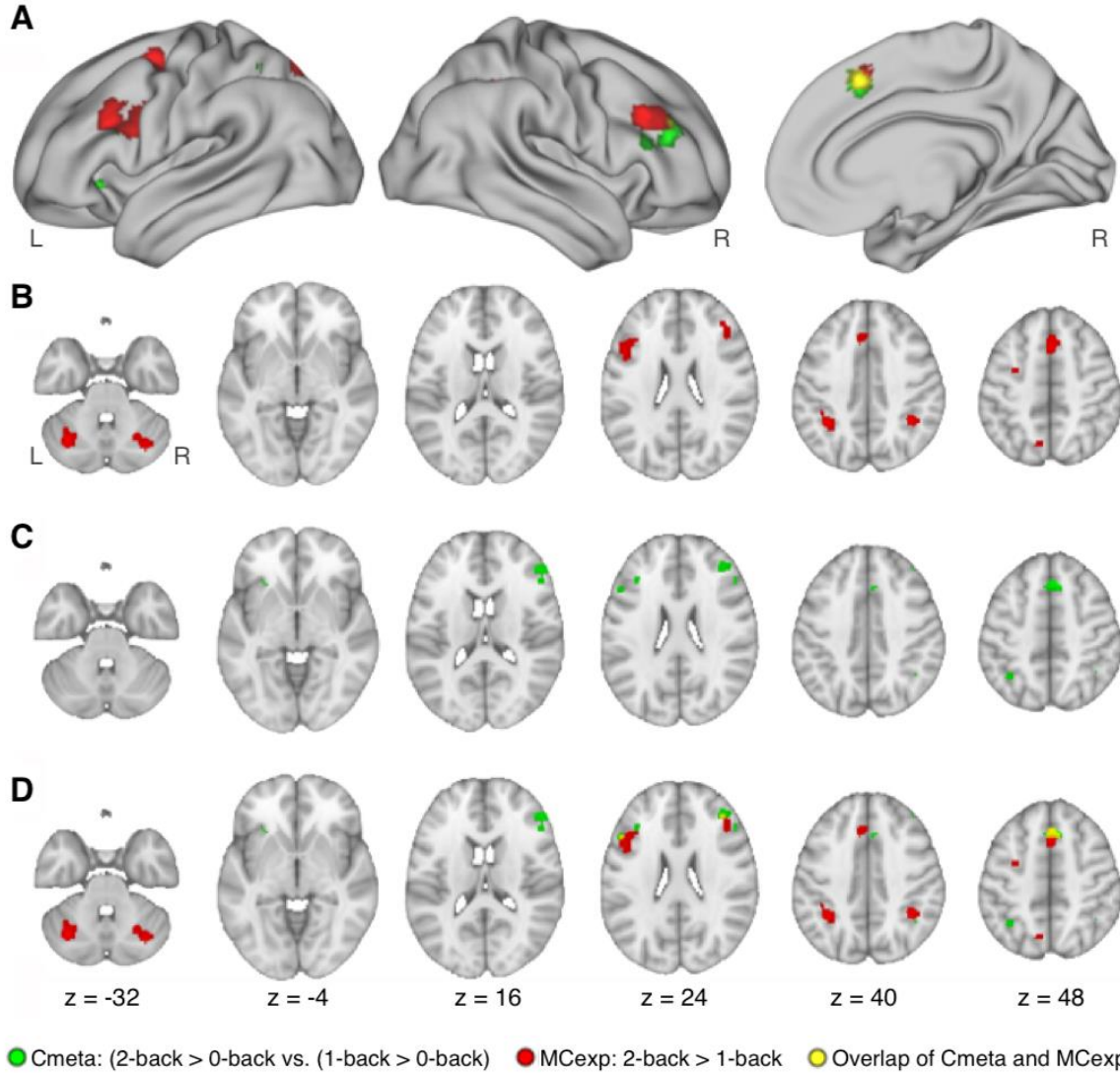


Figure 6. Brain regions revealed by contrast between meta-analyses (2-back > 0-back) > (1-back > 0-back) (Cmeta) and by meta-analyses across 2-back > 1-back experiments (MCexp). (A) and (D) show MCexp in red; Cmeta in green; Overlap of MCexp and Cmeta in yellow. (B) MCexp. (C) Cmeta. Top row, cortex maps. Axial slices in MNI space.

In the contrast analysis comparing 2-back > 0-back with 1-back > 0-back experiments, significant stronger convergence was found in bilateral SFG (medial area), right MFG (dorsal, ventral and ventrolateral area), left MFG (inferior frontal junction), bilateral IFG (area 44), left SPL (lateral, intraparietal area), bilateral IPL (rostradorsal area), left MFG (ventral area) and orbitofrontal gyrus (see Table 3, Figure 6).

Overlapping cluster with the 2-back > 1-back reference network are in the right FP, left IFG (p. Opercularis) and PcG. The 2-back > 1-back reference contrast shows overall a stronger parietal involvement and additional significant convergence in the cerebellum. Although the clusters of both maps are generally in relative proximity to each other, the direct overlap is low (Jaccard-coefficient = 0.051). The computed percent-overlap based

on the BNA 274 ROI-wise comparison shows a much higher degree of similarity (Jaccard-coefficient = 0.441) (Table 2).

This shows that the two Cmetas differ. While the Cmeta with the versus 0-back experiments contains some of the expected areas (i.e. areas of the reference contrast) with significantly stronger convergence, these areas are absent in the Cmeta with the rest/baseline experiments.

Table 3. Brain regions involved in 2-back versus 1-back networks revealed by MCexp and Cmetas

Contrast	Cluster Voxel	Local Peaks (Macroanatomical location)	Cytoarchitectonic Location - Probabilities at local maximum (%)	MNI coordinates			z-score
				x	y	z	
MCexp: 2-back > 1-back	335	L IFG (p. Opercularis) / L MFG	Area 45, 16.5 % Area 44, 15.5 %	-48	18	28	4.62
		L PreCG	Area 44, 26.1 %	-42	6	30	4.3
	229	L LOC, superior division	Area 7A (SPL), 51.1 %	-18	-72	54	4.82
		L LOC, superior division	Area 7A (SPL), 60.7 %	-20	-66	62	3.54
	215	R PcG	Area 6mr/ preSMA, 13.8 %	2	18	48	4.63
	196	L SFG	-	-24	0	56	4.75
	192	R Cerebellum (Crus I)	-	34	-68	-30	4.77
		R Cerebellum Lobule VI	-	28	-62	-28	4.43
	170	L Cerebellum (Crus I)	-	-32	-64	-34	5.26
	158	R MFG	-	42	26	28	4.94
	114	L SMG, posterior division	Area hIP1 (IPS), 55.7 % Area hIP3 (IPS), 23.5 %	-34	-50	38	5.42
	100	R SMG, posterior division	Area hIP1 (IPS), 40.7 % Area hIP2 (IPS), 29.8 % Area hIP3 (IPS), 29.5 %	38	-46	40	4.82
Cmeta: (2-back > rest/ bsl.) > (1-back > rest/ bsl.)	7	L Cerebellum Lobule IV	-	-30	-56	-28	1.76
Cmeta: (2-back > 0-back) > (1-back > 0-back)	210	R FL	-	42	38	18	2.91
		R IFG (p. Opercularis)	Area 45, 63.1 % Area 44, 16.5 %	54	22	28	2.28
	108	R PcG	-	4	26	44	2.7
		R SFG	-	8	24	48	2.68
	34	L IFG (p. Opercularis)	Area 45, 15.3 %	-50	20	26	2.14

32	L SPL	Area hIP3 (IPS), 45.8 % Area 7A (SPL), 27.4 % Area 7PC (SPL), 25.8 %	-34	-54	52	2.23
16	L MFG	-	-34	28	22	2.0
6	R FL/ R MFG	-	38	36	42	1.93
3	L OFC	-	-30	26	-4	1.78
3	R ANG	Area hIP1 (IPS), 41.2 % Area hIP6 (IPS), 6.5%	40	-54	46	1.71
1	R ANG	Area hIP1 (IPS), 41.1 %	42	-54	40	1.7

Note: MCexp = Meta-analysis across contrasts on experimental level, Cmeta = Contrast between two meta-analyses; L = left hemisphere, R = right hemisphere; Cytoarchitectonic areas: Area 44 and Area 45 (K. Amunts et al., 1999; Katrin Amunts et al., 2004); Area 7P (SPL) and Area 7A (SPL) (Scheperjans, Eickhoff, et al., 2008; Scheperjans, Hermann, et al., 2008); Area hIP1 (IPS) and Area hIP2 (IPS) (Choi et al., 2006); Area hIP3 (IPS) (Scheperjans, Eickhoff, et al., 2008; Scheperjans, Hermann, et al., 2008); Area 6mr / preSMA (Ruan et al., 2018).

Evaluation of power in meta-analytical contrasts

To investigate the influence of sample size (i.e. the number of experiments per Cmeta) on the validity of a Cmeta and to verify the previous results, a meta-analysis was simulated using a large sample of subject level contrasts. 8 variations with increasing sample size K (number of studies) were computed and for every variation 100 iterations were calculated. The Cmeta of each iteration was compared with the MCexp within iteration (simulated MCexp) and to the literature derived 2-back > 0-back reference network (literature MCexp).

The following comments on the results are mainly focused on the ROI-wise (based on BNA274) comparisons, as the voxel-wise (whole brain) comparisons are generally lower but show the same trend (depicted in Figure 7 (dashed line) for the sake of completeness). Unless explicitly noted, reported effects refer to ROI-wise comparisons.

Comparison of Cmeta with MCexp (within iteration)

The ROI-wise computed Jaccard coefficient (mean across iterations) between Cmeta and MCexp increases with an increase of K (Figure 7 in blue, solid line). A one-way ANOVA shows that this increase is overall significant, ($F(7, 792) = 82.145$, $p = 2.2e-16$). Post-hoc tests revealed significant increases of similarities ($p < 0.05$) with an increase of 6 studies (K to K+6, i.e. significant increases from 17-23, 20-26, 23-29, 26-32, 29-35, 32-38). The greatest degree of percent-overlap can be observed at K = 38 (Jaccard-coefficient = 0.575 ± 0.033).

The sensitivity increases, analogous to the Jaccard coefficient, with an increase of K (from 0.602 ± 0.061 at $K = 17$ to 0.714 ± 0.043 at $K = 38$). A one-way ANOVA shows a significant effect of sample size on measured sensitivity ($F(7, 792) = 62.162$, $p = 2.2e-16$). Post-hoc tests revealed significant increases in sensitivity ($p < 0.05$) again with an increase of 6 studies (as seen above).

The specificity index on the other hand decreases from 0.901 at $K = 17$ to 0.85 at $K = 38$. Again, one-way ANOVA shows that the effect is significant ($F(7, 792) = 42.222$, $p = 2.2e-16$). In contrast to Jaccard coefficient and sensitivity, post-hoc tests revealed significant decreases of specificity ($p < 0.05$) for the first increases in studies (K), but not between $K = 29$ and any higher K (i.e. significant decreases from 17-20, 20-26, 23-26, 26-38). Noteworthy, voxel-wise measured specificity is almost perfect ($= 1$) and showing almost no decrease between $K = 17$ to $K = 38$ (0.995 to 0.992).

Thus, an increase in K goes along with more regions being correctly but also falsely identified as significant in the Cmeta. However, overall the increase in discovering correct regions seems to outweigh the false discoveries (see Figure 7).

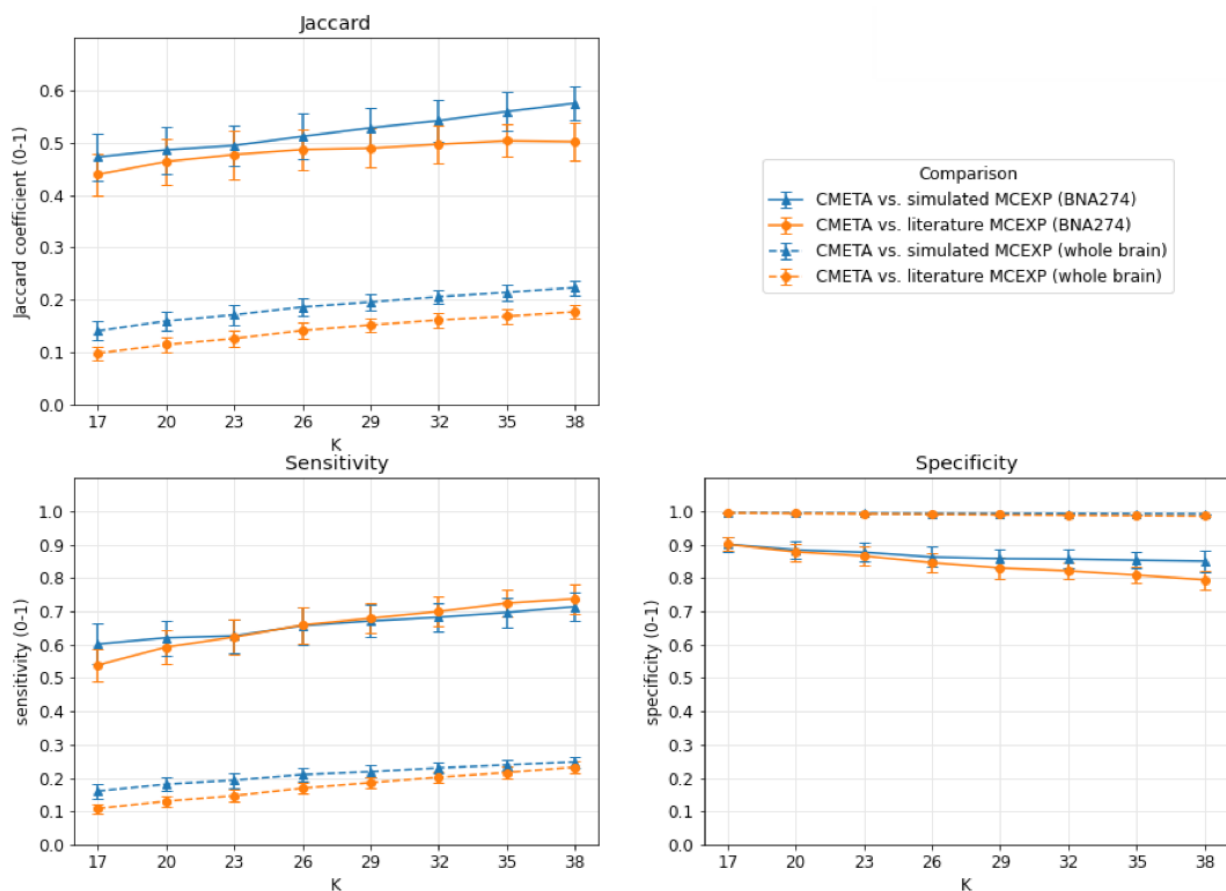


Figure 7. Meta-analytical contrast simulation using a large sample.

Comparison of Cmeta with literature MCexp

The measured similarity between the Cmeta and the literature MCexp shows a similar pattern but overall to a slightly lower degree (Figure 7 in orange, solid line). A one-way ANOVA show overall a significant effect of increasing sample size (K) on the Jaccard-coefficient ($F(7, 792) = 30.724, p = 2.2e-16$). Post-hoc tests revealed significant increases of similarity ($p < 0.05$) only between the first increases in studies (i.e. significant increases from 17-20, 20-26, 23-32). The Jaccard coefficient does not significantly increase between 26 and any higher K, instead it reaches the maximum in similarity at 35 studies (Jaccard-coefficient = 0.503 ± 0.031).

The sensitivity significantly increases (ANOVA: $F(7,792) = 200.92, p = 2.2e-16$) and specificity significantly decreases (ANOVA: $F(7,792) = 179.41, p = 2.2e-16$). Interestingly, both seems to happen at a rate twice as strong compared to the comparison between Cmeta and simulated MCexp (within iteration). The sensitivity increases between $K = 17$ and $K = 38$ from 0.539 to 0.738 and the specificity decreases from 0.901 to 0.795.

Voxel-wise computed specificity is, as in the comparison to the simulated Cmeta, close to 1, with a very small decrease between $K = 17$ to $K = 38$ (0.995 to 0.986).

This indicates that as before an increase in K seems to result in more regions correctly being identified while simultaneously identifying more regions not present in the reference network (see Figure 7 in orange, solid line).

Discussion

The aim of this study was to assess the validity and robustness of meta-analytical contrasts. This was done by comparing a meta-analysis across contrasts on experimental level (MCexp) with a contrast between two meta-analyses (Cmeta), both of which represent the same contrast but are calculated conceptually differently. Next, the influence of the contrast condition in a meta-analytical contrast was examined using the same literature dataset but focusing on a different set of contrasts. Finally, a large dataset was used to simulate differently powered meta-analyses and analyzing the validity as a dependency of number of experiments.

Reference networks: Meta-analysis across contrasts and large-sample n-back contrast

The meta-analysis across 2-back > 0-back experiments (MCexp) shows significant convergence bilaterally in frontal and parietal regions. Those regions are in good accordance with previous WM meta-analytical networks (Owen et al., 2005; Rottschy et al., 2012). However, clusters of the MCexp are slightly smaller compared to the main effect analysis across all WM-tasks by Rottschy and colleagues (2012). A larger sample size (189 vs. 62 experiments) and a more heterogenous sample including a variety of contrasts and WM paradigms, in the WM network, might have led to this difference. In comparison with n-back meta-analysis by Owen et al. (2005), misses the MCexp the right hemispheric pole region. This could likewise be explained by the homogenous sample comparing only the 2-back with the 0-back condition and including mainly verbal paradigms (> 80 %). However, the left-hemispheric lateralization of verbal WM is a debated topic, as the results of a current meta-analysis show (Emch et al., 2019).

The 2-back > 0-back contrast using 435 subject level GLM contrasts shows a similar fronto-parietal pattern of activations as described in the meta-analytical network but is larger in extension. The regions remain mostly the same, with an addition of more frontal involvement, ranging across the MFG. This difference possibly originates from the different samples. The meta-analytical network shows convergence of activations across different groups with slight differences in task designs, pre-processing, analysis and scanning protocol, site. The HCP sample is large but all subjects (age = 22-35) performed the same task, image acquisition and processing/ analysis of data were all standardized (Barch et al., 2013; Glasser et al., 2013; Van Essen et al., 2013). This more homogenous

sample is likely to be the cause for the observed stronger and more extended effects of the contrast.

Meta-analytical contrast evaluation

The whole brain measured Jaccard coefficients between the Cmeta ((2-back > rest/bsl. > (0-back > rest/ bsl.)) and the MCexp and large-sample n-back contrast show a very low degree of similarity (see Table 2). Based on this voxel-wise measure, one could assume that the networks are completely different, the Cmeta not matching the references.

The Jaccard coefficient is commonly used as a reliability metric in fMRI studies to measure the spatial overlap of two whole brain maps (e.g. Kampa et al., 2020; Turner et al., 2018). The coefficient is calculated as a voxel-wise comparison of two thresholded maps. However, this makes the metric dependent on a predefined statistical threshold (Bennett & Miller, 2010). A possibly confounding effect might occur, if the proportion of sign. voxels in both maps differ (Bossier et al., 2020). For example, if two clusters overlap, but in one map the cluster consists of much fewer significant voxels, we would still see a relatively low Jaccard coefficient. In the comparisons performed here, a large difference in the proportion of the number of significant voxels can be seen. Compared to the MCexp network, the Cmeta has 13 times fewer sign. voxels, and compared to the large-sample network, even fewer. However, for the evaluation conducted here, it is not necessarily of interest if every voxel is detected but whether a region is found or not.

In contrast to the voxel-wise measured similarity, a descriptive comparison of cluster locations indicates some accordance between the Cmeta and its reference networks. In the Cmeta significant stronger convergence forming 7 clusters (with 2 consisting of only 3 and 1 voxels) can be observed. All of those clusters lay within or show great overlap to 5 of the 15 clusters of the MCexp. It is striking that the clusters formed in the Cmeta are much smaller than the ones seen in the MCexp. Overlap of both networks can be found in the posterior part of the left SMG (hIP1, hIP2), the right angular gyrus (hIP1), anterior INS (bilateral)/ posterior OFC, right SFG (dorsolateral area 6)/MFG (ventrolateral area 6). However, 10 of the MCexp clusters seem to not be replicated in the Cmeta. The clusters unique to the MCexp include 4 bilateral cerebellar cluster, a cluster spanning between paracingulate gyrus and (pre-) SMA, right Area 45, left SPL, left frontal pole (FL) and left caudate and thalamus. The differences to the large-sample contrast are similar. With the difference that generally a larger part of the brain in the large-sample contrast shows significance compared to the MCexp reference.

To quantify these observed differences and similarities, a comparison was made based on a ROI-wise approach. ROIs are defined by the BNA 274 parcellation (Fan et al., 2016) and declared significant if only one voxel is suprathreshold. This is similar to a cluster-wise comparison by Durnez et al. (2014) declaring a cluster to be replicated if at least one significant voxel is overlapping between images. To make the approach relatively independent of the number and extent of clusters a ROI-wise approach as opposed to a cluster-wise comparison was introduced. The ROI-wise comparison shows a higher degree of overlap between the Cmeta and the reference networks. The measured ROI-wise Jaccard coefficient shows that ca. 27 % of the regions that are significant in the MCexp are also significant in the Cmeta. This includes 22 common regions and 61 regions unique to the MCexp. No regions are unique to the Cmeta (specificity is perfect). The unique regions are similar to the above described cluster locations, including bilateral cerebellar, left thalamus and generally more frontal regions (bilateral IFG, SFG, right frontal pole). The comparison to the large-sample contrast shows generally similar results, though a lower degree of similarity, which can be due to the greater difference in network extent, leading to more regions to be found significant by the ROI-wise measure.

Overall, two common observations can be made from the comparisons. First, all clusters showing significant stronger convergence in the Cmeta are represented in the reference networks. Second, regional differences persist, a considerable amount of brain areas remain undiscovered in the Cmeta. Further, the Cmeta shows considerably smaller clusters compared to the MCexp.

These observations show that the brain regions that show significant stronger convergence in the 2-back > rest/ baseline experiments compared to the 0-back > rest/ baseline experiments are associated with working memory (as observed by the large sample contrast on a study level). Those differences can be interpreted as stronger involvement in one condition (2-back) compared to another condition (0-back), but they seem to represent only a fraction of the brain regions associated with the contrast. Thus, a Cmeta possibly shows only regions which show no (or very low) convergence in the 0-back condition. In contrast to the MCexp which can also reflect smaller differences in the 2-back against 0-back contrast.

Influence of type of contrasting condition

Do Cmeta and MCexp differ dependent on the choice of contrasting condition? To systematically investigate the influence of the contrasting condition on a Cmeta, a second set of meta-analyses was calculated.

The first Cmeta was computed across experiments contrasting against a passive rest/baseline condition, (2-back > rest/ baseline) vs. (1-back > rest/ baseline). A second Cmeta was computed across experiments contrasting against an active control (0-back) condition, (2-back > 0-back) vs. (1-back > 0-back).

The reference network was computed as a MCexp across all experiments reporting the 2-back > 1-back contrast. The 2-back > 1-back meta-analysis, which represents a working memory load effect (Rottschy et al., 2012), shows significant convergence in a similar, though less pronounced fronto-parietal network as the main effect of the 2-back > 0-back meta-analysis. This meta-analytical WM load network is in accordance with previous meta-analyses (Emch et al., 2019; Rottschy et al., 2012). In both previous WM load networks, a more distinct pattern of regions was observed, including more frontal regions. A possible explanation for this weaker 2-back > 1-back network might be the homogenous sample and the relative low sample. Both former studies used a more heterogenous sample of contrasts, including different WM paradigms (i.e. Sternberg, n-back) and different load contrasts (e.g. 3-back > 1-back, modulation by load). In the case Rottschy et al. (2012) the sample was also considerably larger ($n = 44$). Thereby smaller effects might have been revealed in the former but might not be evident in the meta-analysis across 2-back > 1-back.

Due to the reasons stated above, only the ROI-wise comparisons are discussed.

As only on small cerebellar cluster is showing significant stronger convergence in the Cmeta across rest/ baseline experiments, the measured similarities are all low. On the other hand, the Cmeta across control condition experiments shows a rather high degree of similarity with the reference network (Jaccard-coefficient = 0.441). This is also higher than similarity observed in the first meta-analytical contrast evaluation (compare Table 2).

The high degree of sensitivity (TPR = 0.517) of the Cmeta across control conditions comes with a decrease of specificity slightly below 1. This becomes evident through a descriptive comparison. A cluster in the left MFG (ventral area) and a cluster in the left OFG seem to

be uniquely identified in the Cmeta. It should be further noted, that most clusters are not truly overlapping between maps but are in relatively proximity to each other. It therefore seems that the choice of contrasting condition directly effects Cmetas.

Early studies highlighted the implications that different contrasting conditions entail (Price et al., 1997; Raichle, 1998) and found resting-state measurements to be a suboptimal baseline condition as it potentially alters the activity during the task condition (Stark & Squire, 2001). A meta-analysis by Price et al. (2005) supports those findings, showing that a contrast with a high-level baseline has a higher sensitivity to the cognitive processes of interest compared to a low-level baseline contrast. Those findings fit nicely with the here presented results. High-level contrasts (in this case 0-back) might provide generally higher sensitivity and thus show more convergence across experiments on a meta-analytical level. However, it is important to emphasize that this does not necessarily imply that the effect of interest is generally not observable in meta-analyses across low-level baseline experiments. But might rather suggest that an effect was too weak to be captured by the Cmeta (across rest/ baseline experiments).

The observed reduced specificity in the Cmeta across control experiments may be due to two reasons. First, the Cmeta might show significant stronger convergence in regions not associated with the contrasted cognitive processes (i.e. false positives). However, this seems to be relatively unlikely as all identified regions are within previously reported WM load effect networks (Emch et al., 2019; Rottschy et al., 2012). Second, the MCexp reference network might miss some of the regions associated with the process of interest (i.e. be incomplete). This is possibly because the sample is too small ($N = 20$) not showing convergence in all regions associated with WM load. Further, some studies reported the 2-back > 1-back contrast to be not significant, which potentially indicates a generally unreliable contrast at study level (Esteves et al., 2018; Migo et al., 2015; Sapara et al., 2014).

Overall, these findings suggest that the contrasting conditions in experiments seem to have a big influence on the results of a contrast between two meta-analyses. That is, experiments with a contrast against a rest/ baseline condition seem to represent the process of interest to a weaker degree compared to experiments against a more active control condition. Thus, potentially confounding meta-analytical results by showing no/low convergence in relevant areas. Further investigations are highly advisable as this might be a source of strong bias when performing a Cmeta.

Are experiments contrasting vs. rest suited for a CBMA?

Contrasts versus a resting state/ baseline condition should generally show more activations as the same contrast versus a control condition (Barch et al., 2013). This is due to the inherent subtraction logic of contrast analysis (Price et al., 1997).

Therefore unexpected was the observed size of the 2-back > rest/ baseline meta-analytical network, which is smaller than the 2-back > 0-back network (compare Supplementary Figure 1 and Figure 3). Working memory processes are theoretically isolated by contrasting the 2-back versus the 0-back condition (Barch et al., 2013; Miller et al., 2009). Thus, the 2-back > 0-back network should show only the regions involved in working memory. It would be expected, that the 2-back > rest/baseline network includes all those regions and additionally with low-level task demand associated regions (i.e. motor, visual, attention). In contrast to these expectations, the meta-analysis across 2-back > rest/ baseline experiments shows only significant convergence in regions within the 2-back > 0-back reference network and not beyond.

A first explanation for this might be the issue of power ($N = 31$ vs. $N = 62$). However, this does not seem to be the case, as shown by the results of a second version of the meta-analysis across 2-back > 0-back experiments with a reduced sample size (compare Supplementary Figure 3). This reduced version when compared to the full sample network, shows a very similar pattern of significant convergence with generally slightly smaller clusters. It only misses a small left parietal and the thalamus clusters. It however is still clearly bigger and shows more regional convergence compared to the 2-back > rest network.

These observations, together with the results from the Cmeta contrasting condition investigation, lead to the general question if contrasts against a fixation/ rest condition are suited for a coordinate based meta-analysis (CBMA). In this specific case, if ALE meta-analysis is suited to accurately capture the entirety of the cognitive processes involved in the here investigated contrasts.

Coordinate-based meta-analyses are known for a loss of information and not capturing smaller effect sizes (Salimi-Khorshidi et al., 2009). The basis of a CBMA are the reported foci or peak coordinates of clusters in published studies. The coordinates are serving as a proxy for the location of the found activation differences. Their high availability is what makes CBMA so attractive but using only the coordinates comes with some costs.

Information about the cluster size (extent, covered locations) and z-score are not included in the ALE algorithm (Eickhoff et al., 2009). Usually 10 foci are reported per experiment (median of 161 peer-reviewed experiments from the here used n-back dataset). The number of coordinates reported per cluster are usually between 1-3.

This loss of information may be a problem in some cases. This might be in the case in contrasts versus a rest/ fixation baseline, especially higher cognitive contrasts (e.g. 2-back > fixation) and in studies with a large sample size. In both cases, the clusters become larger. For contrasts with a vs. rest control condition, because multiple processes are covered by the contrast. With larger sample, clusters also become larger as additional smaller effects are revealed. Larger clusters may eventually merge. In the most extreme case, only one large contiguous cluster can then be observed (Bossier et al., 2020). An example of a contrast where only one large cluster was observed in a relatively large sample is by Aguilar-Ortiz and colleagues (2020). They found a large contiguous cluster of 115980 voxels in a sample of $N = 67$ for the contrast 2-back > 1-back. Another example of a contrast against a rest/fixation baseline in which only a single cluster was observed in a smaller sample ($N = 26$) is from Rodríguez-Cano and colleagues (2017). They found a large contiguous cluster with the size of 71142 voxels for the contrast 2-back > baseline. These large clusters are clearly insufficiently described by a single coordinate.

The problem of large clusters also occurred in the sent contrasts. In the results kindly provided by Prof. Dr. Jacob Lahr and Dr. Lora Minkova (Lahr et al., 2018), single large clusters per contrast can be observed. Two other examples of 2-back > rest/fixation contrasts where large clusters can be observed are those kindly provided by Dr. Ian Harding (Harding et al., 2016) and Dr. Yu Fukuda (Fukuda et al., 2019). In one case, only two large clusters are observed with 43803 and 4291 voxels with a sample size of $N = 34$. In the other case, among others, a large cluster with 55711 voxels is found ($N = 24$). If only the first 3 peak coordinates of such large clusters are included in the meta-analysis, probably only a fraction of the real effects are covered. From the received results, the first 10 peaks per cluster were extracted if possible. This was possible if image files were provided. The standard local maxima output table of SPM, however, reports only 3 peaks per cluster. But even 10 peaks are probably not sufficient to include everything in consideration of such large clusters.

The observations might imply that some contrasts may not be ideal for an ALE meta-analysis. In the meta-analytical contrast studied here, this may have resulted in a bias.

More studies alone are not enough for a robust meta-analytical contrast

To what extent is the validity of Cmetas driven by the number of experiments? To assess this question large-sample-simulated meta-analyses were conducted.

As observed in the literature-based meta-analysis, the ROI-wise calculated similarities show considerably better results compared to the voxel-wise measures. For the reasons discussed above, a ROI-wise metric might be the preferred choice to assess the similarity between Cmeta and reference networks, as it is predominantly of interest whether regions are found. Therefore, mainly the results of this approach will be discussed.

The comparison of the Cmeta with the MCexp within the same iteration (i.e. both computed from the same set of studies) seem to generally show a higher degree of similarity as the comparison with the literature MCexp (compare Figure 7). This might be due to an increase of power in the MCexp with an increase in included studies (K). Higher powered meta-analyses might detect smaller/ medium sized effects (Eickhoff et al., 2016) and lead to more and larger cluster, which will lead to more associated ROIs with an increase of K.

In general, there is an increase with the number of K in similarity (measured by Jaccard-coefficient) and sensitivity and a decrease of specificity. Though, even a relatively high power (K = 38) seems to not capture the full network as seen in the reference contrasts. The maximum reached similarity is 0.575 (at K = 38, Cmeta vs. within MCexp). The highest observed sensitivity is 0.738 (at K = 38, Cmeta vs. literature MCexp). Thus, even computing a contrast between two very robust meta-analyses (38 experiments each), approximately 26 % of all relevant regions are still not detected and 15-20 % are falsely identified (as indicated by specificity of 0.795 and 0.85 at K = 38).

While increasing the sample size is shown to be one solution to replicability in task fMRI group-level contrasts (Bossier et al., 2020; Cremers et al., 2017; Geuter et al., 2018), it seems that an increase in sample size is only showing a relatively moderate positive effect on the validity of Cmetas. Thus, it seems to be not the only factor towards robust Cmetas. Further considerations and potential confounds are in detail discussed in the next chapters.

How many experiments are recommendable for a Cmeta?

Generally, conducting meta-analyses follows the wisdom “take what you get”, i.e. including all available experiments. However, it is important to know whether a particular

hypothesis should be tested based on the available sample (Müller et al., 2018). So, whether one should calculate a Cmeta if the available number of experiments is relatively small; or whether it would be better to loosen the inclusion criteria to increase power; or whether a very high number of experiments could possibly introduce errors.

The results of the large-sample-simulated meta-analysis suggest that there is a trade of between sensitivity and specificity. The comparison to the literature reference network shows no significant increase in similarity from 26 to any higher K. It seems to be the balance point where sensitivity and specificity increase and decrease at approximately similar rates.

The decrease in specificity shows an inflation of false positives. A way to control them could be to correct the Cmeta results (results here are uncorrected) (Eickhoff et al., 2011; Laird, Fox, et al., 2005). However, Cmeta correction might also come with a decrease in sensitivity. Although this study cannot make any assumptions on this, it might improve the problem.

Whole brain, voxel-wise comparisons of the Cmeta with reference networks shows nearly perfect specificity (compare Figure 7). This might indicate that the observed specificity decrease is a problem introduced through the ROI-wise assessment of regional overlap. Thus, a verification of the current method and exploration of other methods should be performed.

A higher number of experiments goes along with a higher sensitivity. However, it is important to considerate that larger samples also increase the likelihood of false positives. That said, if strong confidence in the results is essential than a sample of 26 experiments should not be exceeded by too much. If including more experiments, it is advisable to use a more conservative threshold (i.e. p-value) or possibly correct for multiple comparisons.

Which effects can be shown in a Cmeta?

The Cmeta is computed as a statistical comparison of two meta-analyses. And the ALE maps of those meta-analyses are computed on the basis of peak activation coordinates. Thus, a Cmeta is a comparison of activations. However, a contrast from individual fMRI studies is usually computed between activations and deactivations. A contrast (e.g. 2-back > 0-back) can be masked with the relevant main effect (e.g. 2-back) to ensure that regions with relative deactivations in one condition (e.g. 0-back) and no activations in the other

condition (e.g. 2-back) don't lead to positive results (Schlösser et al., 2007). Assumed that this masking procedure is rarely done (not reported once in the literature sample), it is possible that some regions showing convergence in the literature MCexp network are due to deactivations in the control task. However, the Cmeta will not show significant stronger convergence in those regions, as deactivations are not included, and those regions are not showing activations in the condition of interest. Thus, some measured false negatives in the Cmeta might actually not be false. To ensure that the reference network is only showing convergence across "real" activations only masked contrasts should be fed in the meta-analysis.

How strong needs a difference in convergence be, in order to be found significant?

The individual 2-back > rest/ bsl. and 0-back > rest/bsl. meta-analyses (from the first the meta-analytical contrast investigation) both show convergence in the bilateral aINS and (pre-)SMA (see Supplementary Figure 1). This is expected as both regions have been associated with WM (Rottschy et al., 2012) and attention (Langner & Eickhoff, 2013). Interestingly, the Cmeta between both meta-analyses finds significant stronger convergence in the aINS but not in the bilateral (pre-)SMA, while both areas show convergence in the MCexp reference network. The significantly stronger convergence in the aINS shows that the Cmeta is able to find a difference in convergence even in regions that show convergence in both meta-analyses. This is in line with observation that can be made in the Cmeta between task-set versus task-load effects by Rottschy et al. (2012). The meta-analysis shows significant stronger convergence in load effects in frontal areas that are also significant in a conjunction analysis between both meta-analyses. However, the absence of the (pre-)SMA raises the question of how strong a difference in convergence must be in order to be identified?

The first example illustrates the conceptual differences between Cmeta and MCexp and raises the question if perfect sensitivity is theoretically possible or whether the two contrasts are different after all. The second case shows that a Cmeta is potentially able to detect differences in convergence, even if this region is in both meta-analyses significant.

General discussion

A Cmeta is not a substitute for contrasts on experimental level because not all relevant brain regions are identified. The preferred choice should always be to perform a MCexp. In cases where this is not feasible due to a limited number of studies, Cmetas are still

informative as the regions identified are most likely true positives. Nevertheless, it should be kept in mind that never all regions are identified, i.e. the absence of regions in Cmetas should be interpreted cautiously or omitted.

There are a number of possible reasons why not all regions are found. It may be a problem at the level of individual experiments, but also on the meta-analytical level. Regarding the former a problem might be the coordinates that are fed into the meta-analysis, the contrasted conditions, (no) correction for multiple comparisons, the number of coordinates reported, and the number of subjects per experiment. Another possibility is that the problem arises at the meta-analytical level. For example, the number of experiments per meta-analysis or possibly an imbalanced design could lead to a bias. Moreover, in the calculation, the Cmeta is masked with the meta-analysis main effect and not corrected for multiple comparison.

Potential problems on the experimental level

Higher level contrasts versus a fixation baseline may not be ideal for a neuroimaging meta-analysis. This is also true for large sample contrasts. As discussed, the problem could be a merging of clusters, resulting in larger and fewer clusters that may not be adequately described by peak coordinates (foci). To conclude, if the meta-analyses cannot detect all relevant regions due to inaccurate / too little information in the coordinates (as evidenced by smaller networks), then the Cmeta certainly cannot.

Related to this, the choice of a threshold for group-level maps and whether to correct for multiple comparisons affects the number and size of clusters (Bossier et al., 2020) and thus indirectly affects the coordinates reported. While corrected results might provide smaller cluster, uncorrected results might provide more cluster. However, in a scenario where all peaks are reported, uncorrected results might be more advantageous. That is, because the ALE algorithm eventually separates random occurring clustering (across experiments) from significant convergence (Eickhoff et al., 2009). In addition, the choice of the number of peak coordinates extracted can also be a critical factor. If only the first 3 peak coordinates of every cluster are extracted, it is likely that big cluster with multiple subpeaks will be inadequately represented.

As the potential influence of those factors ((un)corrected sample size, number of extracted peak coordinates and number of subjects per study) on the Cmeta results is still unknown, a systemic investigation is advisable.

Potential problems on the meta-analytical level

A central limitation to all the literature-based assessments was sample size. While above the for meta-analyses recommended 17-20 experiments (Eickhoff et al., 2016), the individual samples might still be not sufficiently large to get robust findings. The fact that many regions are not detected in the Cmeta that show significant convergence in MCexp suggests a power problem. The results of the large-sample-simulation meta-analysis partially contradict these assumptions. There is a positive effect of power on the number of regions detected in the Cmeta, but this is relatively weak and only partially explains the differences to the MCexp.

Another important aspect concerns unequal sample sizes. As seen in the current study, the number of experiments can vary widely. When calculating a contrast between two very unequal meta-analyses, this could introduce another potential bias. The chosen method to reduce the number of experiments of the meta-analysis with the higher number by matching the number of experiments manually could be prone to subjective error and may not be the ideal approach. Therefore, a subsampling method might be better to correct for the effect of unequal samples (Gu et al., 2019; Poudel et al., 2020).

Masking and multiple comparison correction in meta-analytical contrast

Besides possible confounds caused by the extracted coordinates or the study design, the calculation of the Cmeta may be biased.

After calculating the contrast between two ALE meta-analyses, it is masked with the result of the corresponding main effect analysis (Eickhoff et al., 2011). In the contrast analysis between the 2-back > rest/baseline and 0-back > rest/baseline meta-analyses, the calculated contrast map was masked with the results of the 2-back > rest/baseline main effect analysis. In this case, only one-sided, because only the 2-back > 0-back (not 0-back > 2-back) contrast was of interest. This is similar to masking in individual fMRI studies (Schlösser et al., 2007), where a contrast is masked with the main effect to facilitate interpretation and to obtain only differences that generally show convergence in the respective condition (as described above in detail). However, it could be investigated if this step is necessary or could be omitted, to include more potential regions.

Correction of multiple comparisons is necessary when a statistical test is repeated several times, thereby increasing the probability of a false positive result (type I error) (Shaffer, 1995). The results of the literature-based Cmeta do not show false positives, but rather a

problem of false negatives. This seems more likely to be due to insufficient power to identify all effects. In the case of the meta-analysis simulated with the large data set, there are significantly fewer false negatives, but an increase in false positives with an increase in power. This suggests that testing for multiple comparisons is likely to be useful, at least above a certain number of experiments.

Limitations

While the simulated approach can be controlled for a lot of influences it might help to uncover the importance of single factors on Cmetas. This means the results are likely to show high internal validity (Onwuegbuzie, 2000). This comes at the expense that it is unclear how realistic and generalizable the results are for other, literature-based meta-analyses. Due to this low external validity, the implications made from this approach should be critically weighed. The literature (traditional) meta-analysis approach is based on a heterogeneous sample of studies. Therefore, the external validity can be regarded relatively high. However due to the limitation in availability of data it remains unclear to what extent single factors might have influenced the results (low internal validity).

An ALE meta-analysis is limited by the existing literature. This was particularly evident in the varying number of experiments found for each contrast. There may be some publication bias associated with the literature (Acar et al., 2018), which, as mentioned earlier, might manifest itself more strongly for certain contrasts. The available literature results were extended by several authors who contributed their additional results upon request. However, this also means that the results are not peer-reviewed, leading to another possible confound (Müller et al., 2018).

The resulting dataset remained suboptimal in two aspects. It contained only the just recommended number of experiments (17-20) for some meta-analyses and all three COIs (*condition A > condition B*; *condition A > baseline*; *condition B > baseline*) were only available for some studies. For the 2-back > 0-back comparison, all three experiments/contrasts of interest (0-back > rest; 2-back > rest; 2-back > 0-back) were retrieved for only 10 subject groups. This restricts the comparison to between-studies effects instead of comparing within-studies effects.

The large-sample-simulated meta-analysis approach is limited in various ways. Firstly, in its attempt to create a realistic scenario of an actual CBMA. Some factors were varied while others were fixed. Only one analysis pipeline was used, opposed to the analytical

flexibility observed in the literature (Carp, 2012). The used HCP sample (Van Essen et al., 2013) is very homogenous in terms of subject age, scanner protocol and overall of high quality, thereby not necessarily reflecting a typical fMRI study. Some variability was introduced by drawing differently sized studies and varying the group map thresholding/correction method. However, this could also have biased the results, as discussed previously.

Both approaches are based on the same principle, a contrast comparison of one specific cognitive task contrast. It remains open how universal potential findings are, in respect to other contrasts (e.g. other tasks, groups etc.).

Lastly, all discussed results are based on the measure of similarity. The here proposed qualitative assessment using a ROI-wise comparison is based on no threshold. This very liberal decision, declaring parcels as significant by just 1 voxel, might have inflated false declarations. This could be addressed by choosing a threshold (i.e. a minimum of N voxels or a minimum of X percent of voxels found to be significant). This was not done to no run into the same problem discussed prior, a similarity difference based on the cluster extent. Further, any chosen threshold would be arbitrary. In order not to eventually cancel out true effects, no threshold was chosen, to the cost that this may falsely label regions as significant. A further important consideration for this similarity measure is the classification of ROIs. Using a parcellation, it is crucial to assume that the ROIs are representing meaningful brain areas. Therefore the Brainnetome Atlas was chosen, representing parcels based on known anatomical and functional connections (Fan et al., 2016). It is clearly essential to have a valid tool of assessing similarities. Therefore, is a validation of the measure and exploration of further metrics highly advisable.

Outlook

As described in detail in the “General discussion” section, there are many potential sources of confounds for a meta-analytical contrast. Several investigations are necessary to rule them out and to give a more thorough recommendation for the use of Cmetas.

To confirm the results of the literature-based analysis, a "complete" literature sample (i.e. all 3 contrasts for at least 20 studies) would be needed. This would change the comparison from between-study effects to within-study effects. In addition, other cognitive tasks and possibly group activation differences could be compared on a meta-analytical level. This is important to rule out the possibility that the chosen contrast was not suitable or that

certain contrasts (higher cognitive functions versus rest baseline) are generally unsuitable for CBMA.

In the light of the large-sample-simulated meta-analysis, further research should be conducted to test different properties. At the study level, the influence of the chosen threshold method and the influence of the number of coordinates remained unclear. Variations with only one threshold method (corrected vs. uncorrected) and selecting all coordinates or only the first 3 peak coordinates could provide further insights. Also, the effect of unequal sample sizes in the experiments seemed to play a crucial role in understanding Cmetas. Variations of study samples using only a steady number of subjects per study could help to investigate this factor. Eventually, all comparisons should be replicated in different samples and with different tasks to validate the results.

Another insightful investigation would test the influence of deactivations on Cmetas. The same large-sample-simulated meta-analysis approach could be used, but with a slightly different reference MCexp. Instead of extracting the coordinates directly from the 2-back > 0-back group analyses, they could be masked beforehand with the 2-back > fixation contrast. This would ensure that any positive effects observed in the 2-back > 0-back contrasts were due to activation differences and not deactivations in the 0-back condition.

Finally, the current algorithm should be re-evaluated and finally optimized. A subsampled Cmeta calculation should be investigated to test the effect of imbalanced designs. The effect of masking and lack of multiple comparison correction in Cmetas remains unclear and further investigation would clearly be advisable.

Conclusion

The results indicate that although a contrast between two meta-analyses cannot necessarily identify all regions, the regions found can be interpreted with relative confidence similar to regions found in a meta-analysis across experiments on experimental level (i.e. probability of a false positive is low). However, it is not recommended to interpret the absence of regions, because of the observed low sensitivity.

Experiments contrasting against rest/ baseline condition seem to be less suitable than experiments contrasting against control condition, i.e. the results recommend using a high-level control if possible.

Finally, an increase in the number of cases is associated with a better sensitivity and similarity, but it seems that from a number of 26 experiments onward this increase is reduced with a higher likelihood of seeing false positives.

Bibliography

- Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, 8, 14.
<https://doi.org/10.3389/fninf.2014.00014>
- Acar, F., Seurinck, R., Eickhoff, S. B., & Moerkerke, B. (2018). Assessing robustness against potential publication bias in Activation Likelihood Estimation (ALE) meta-analyses for fMRI. *PLOS ONE*, 13(11), e0208177.
<https://doi.org/10.1371/journal.pone.0208177>
- Aguilar-Ortiz, S., Salgado-Pineda, P., Vega, D., Pascual, J. C., Marco-Pallarés, J., Soler, J., Brunel, C., Martin-Blanco, A., Soto, A., Ribas, J., Maristany, T., Sarró, S., Rodríguez-Fornells, A., Salvador, R., McKenna, P. J., & Pomarol-Clotet, E. (2020). Evidence for default mode network dysfunction in borderline personality disorder. *Psychological Medicine*, 50(10), 1746–1754.
<https://doi.org/10.1017/S0033291719001880>
- Altman, D. G., & Bland, J. M. (1994). Statistics Notes: Diagnostic tests 1: sensitivity and specificity. *BMJ*, 308(6943), 1552–1552.
<https://doi.org/10.1136/bmj.308.6943.1552>
- Amunts, K., Schleicher, A., Bürgel, U., Mohlberg, H., Uylings, H. B., & Zilles, K. (1999). Broca's region revisited: Cytoarchitecture and intersubject variability. *The Journal of Comparative Neurology*, 412(2), 319–341.
[https://doi.org/10.1002/\(sici\)1096-9861\(19990920\)412:2<319::aid-cne10>3.0.co;2-7](https://doi.org/10.1002/(sici)1096-9861(19990920)412:2<319::aid-cne10>3.0.co;2-7)
- Amunts, Katrin, Weiss, P. H., Mohlberg, H., Pieperhoff, P., Eickhoff, S., Gurd, J. M., Marshall, J. C., Shah, N. J., Fink, G. R., & Zilles, K. (2004). Analysis of neural mechanisms underlying verbal fluency in cytoarchitectonically defined stereotaxic space—The roles of Brodmann areas 44 and 45. *NeuroImage*, 22(1), 42–56. <https://doi.org/10.1016/j.neuroimage.2003.12.031>
- Barch, D. M., Burgess, G. C., Harms, M. P., Petersen, S. E., Schlaggar, B. L., Corbetta, M., Glasser, M. F., Curtiss, S., Dixit, S., Feldt, C., Nolan, D., Bryant, E., Hartley, T., Footer, O., Bjork, J. M., Poldrack, R., Smith, S., Johansen-Berg, H., Snyder, A. Z., & Van Essen, D. C. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *NeuroImage*, 80, 169–189.
<https://doi.org/10.1016/j.neuroimage.2013.05.033>
- Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences*, 1191, 133–155. <https://doi.org/10.1111/j.1749-6632.2010.05446.x>
- Blokland, G. A. M., McMahon, K. L., Thompson, P. M., Martin, N. G., de Zubicaray, G. I., & Wright, M. J. (2011). Heritability of Working Memory Brain Activation. *Journal of Neuroscience*, 31(30), 10882–10890. <https://doi.org/10.1523/JNEUROSCI.5334-10.2011>
- Bossier, H., Roels, S. P., Seurinck, R., Banaschewski, T., Barker, G. J., Bokde, A. L. W., Quinlan, E. B., Desrivieres, S., Flor, H., Grigis, A., Garavan, H., Gowland, P., Heinz, A.,

- Ittermann, B., Martinot, J.-L., Artiges, E., Nees, F., Orfanos, D. P., Poustka, L., ... Moerkerke, B. (2020). The empirical replicability of task-based fMRI as a function of sample size. *NeuroImage*, 212, 116601. <https://doi.org/10.1016/j.neuroimage.2020.116601>
- Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., ... Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810), 84–88. <https://doi.org/10.1038/s41586-020-2314-9>
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365–376. <https://doi.org/10.1038/nrn3475>
- Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A. R., Langner, R., & Eickhoff, S. B. (2012). Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Structure and Function*, 217(4), 783–796. <https://doi.org/10.1007/s00429-012-0380-y>
- Carp, J. (2012). On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments. *Frontiers in Neuroscience*, 6. <https://doi.org/10.3389/fnins.2012.00149>
- Choi, H.-J., Zilles, K., Mohlberg, H., Schleicher, A., Fink, G. R., Armstrong, E., & Amunts, K. (2006). Cytoarchitectonic identification and probabilistic mapping of two distinct areas within the anterior ventral bank of the human intraparietal sulcus. *The Journal of Comparative Neurology*, 495(1), 53–69. <https://doi.org/10.1002/cne.20849>
- Chuan-Peng, H., Huang, Y., Eickhoff, S. B., Peng, K., & Sui, J. (2020). Seeking the “Beauty Center” in the Brain: A Meta-Analysis of fMRI Studies of Beautiful Human Faces and Visual Art. *Cognitive, Affective, & Behavioral Neuroscience*. <https://doi.org/10.3758/s13415-020-00827-z>
- Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Elsevier Science. <https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=923159>
- Cremers, H. R., Wager, T. D., & Yarkoni, T. (2017). The relation between statistical power and inference in fMRI. *PLOS ONE*, 12(11), e0184923. <https://doi.org/10.1371/journal.pone.0184923>
- Desmond, J. E., & Glover, G. H. (2002). Estimating sample size in functional MRI (fMRI) neuroimaging studies: Statistical power analyses. *Journal of Neuroscience Methods*, 118(2), 115–128. [https://doi.org/10.1016/S0165-0270\(02\)00121-8](https://doi.org/10.1016/S0165-0270(02)00121-8)
- Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3), 297–302. <https://doi.org/10.2307/1932409>
- Diedrichsen, J., Maderwald, S., Küper, M., Thürling, M., Rabe, K., Gizewski, E. R., Ladd, M. E., & Timmann, D. (2011). Imaging the deep cerebellar nuclei: A probabilistic

- atlas and normalization procedure. *NeuroImage*, 54(3), 1786–1794.
<https://doi.org/10.1016/j.neuroimage.2010.10.035>
- Diedrichsen, Jörn, Balsters, J. H., Flavell, J., Cussans, E., & Ramnani, N. (2009). A probabilistic MR atlas of the human cerebellum. *NeuroImage*, 46(1), 39–46.
<https://doi.org/10.1016/j.neuroimage.2009.01.045>
- Durnez, J., Moerkerke, B., & Nichols, T. E. (2014). Post-hoc power estimation for topological inference in fMRI. *NeuroImage*, 84, 45–64.
<https://doi.org/10.1016/j.neuroimage.2013.07.072>
- Eickhoff, S. B., Bzdok, D., Laird, A. R., Kurth, F., & Fox, P. T. (2012). Activation likelihood estimation meta-analysis revisited. *NeuroImage*, 59(3), 2349–2361.
<https://doi.org/10.1016/j.neuroimage.2011.09.017>
- Eickhoff, S. B., Bzdok, D., Laird, A. R., Roski, C., Caspers, S., Zilles, K., & Fox, P. T. (2011). Co-activation patterns distinguish cortical modules, their connectivity and functional differentiation. *NeuroImage*, 57(3), 938–949.
<https://doi.org/10.1016/j.neuroimage.2011.05.021>
- Eickhoff, S. B., Laird, A. R., Grefkes, C., Wang, L. E., Zilles, K., & Fox, P. T. (2009). Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: A random-effects approach based on empirical estimates of spatial uncertainty. *Human Brain Mapping*, 30(9), 2907–2926.
<https://doi.org/10.1002/hbm.20718>
- Eickhoff, S. B., Nichols, T. E., Laird, A. R., Hoffstaedter, F., Amunts, K., Fox, P. T., Bzdok, D., & Eickhoff, C. R. (2016). Behavior, sensitivity, and power of activation likelihood estimation characterized by massive empirical simulation. *NeuroImage*, 137, 70–85. <https://doi.org/10.1016/j.neuroimage.2016.04.072>
- Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage*, 25(4), 1325–1335.
<https://doi.org/10.1016/j.neuroimage.2004.12.034>
- Emch, M., von Bastian, C. C., & Koch, K. (2019). Neural Correlates of Verbal Working Memory: An fMRI Meta-Analysis. *Frontiers in Human Neuroscience*, 13, 180.
<https://doi.org/10.3389/fnhum.2019.00180>
- Esteban, O., Ciric, R., Finc, K., Durnez, J., Ghosh, S., & Poldrack, R. A. (2019). *Task fMRI Analysis using FSL and data preprocessed with fMRIPrep*. Zenodo.
<https://doi.org/10.5281/zenodo.2635848>
- Esteban, O., Markiewicz, C. J., Burns, C., Goncalves, M., Jarecka, D., Ziegler, E., Berleant, S., Ellis, D. G., Pinsard, B., Madison, C., Waskom, M., Notter, M. P., Clark, D., Manhães-Savio, A., Clark, D., Jordan, K., Dayan, M., Halchenko, Y. O., Loney, F., ... Ghosh, S. (2020). *nipy/nipype: 1.5.0*. Zenodo. <https://doi.org/10.5281/zenodo.3874968>
- Esteves, M., Magalhães, R., Marques, P., Castanho, T. C., Portugal-Nunes, C., Soares, J. M., Almeida, A., Santos, N. C., Sousa, N., & Leite-Almeida, H. (2018). Functional Hemispheric (A)symmetries in the Aged Brain—Relevance for Working Memory. *Frontiers in Aging Neuroscience*, 10, 58.
<https://doi.org/10.3389/fnagi.2018.00058>

- Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A. R., Fox, P. T., Eickhoff, S. B., Yu, C., & Jiang, T. (2016). The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cerebral Cortex (New York, N.Y.: 1991)*, 26(8), 3508–3526. <https://doi.org/10.1093/cercor/bhw157>
- Ferstl, E. C., Neumann, J., Bogler, C., & von Cramon, D. Y. (2008). The extended language network: A meta-analysis of neuroimaging studies on text comprehension. *Human Brain Mapping*, 29(5), 581–593. <https://doi.org/10.1002/hbm.20422>
- Fukuda, Y., Katthagen, T., Deserno, L., Shayegan, L., Kaminski, J., Heinz, A., & Schlagenhauf, F. (2019). Reduced parietofrontal effective connectivity during a working-memory task in people with high delusional ideation. *Journal of Psychiatry & Neuroscience*, 44(3), 195–204. <https://doi.org/10.1503/jpn.180043>
- Geuter, S., Qi, G., Welsh, R. C., Wager, T. D., & Lindquist, M. A. (2018). *Effect Size and Power in fMRI Group Analysis* [Preprint]. Neuroscience. <https://doi.org/10.1101/295048>
- Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., & Jenkinson, M. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, 80, 105–124. <https://doi.org/10.1016/j.neuroimage.2013.04.127>
- Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, 48(1), 63–72. <https://doi.org/10.1016/j.neuroimage.2009.06.060>
- Gu, R., Huang, W., Camilleri, J., Xu, P., Wei, P., Eickhoff, S. B., & Feng, C. (2019). Love is analogous to money in human brain: Coordinate-based and functional connectivity meta-analyses of social and monetary reward anticipation. *Neuroscience & Biobehavioral Reviews*, 100, 108–128. <https://doi.org/10.1016/j.neubiorev.2019.02.017>
- Harding, I. H., Corben, L. A., Storey, E., Egan, G. F., Stagnitti, M. R., Poudel, G. R., Delatycki, M. B., & Georgiou-Karistianis, N. (2016). Fronto-cerebellar dysfunction and dysconnectivity underlying cognition in friedreich ataxia: The IMAGE-FRDA study: Cognitive Networks in Friedreich Ataxia. *Human Brain Mapping*, 37(1), 338–350. <https://doi.org/10.1002/hbm.23034>
- HCP Data Release Updates: Known Issues and Planned fixes—Connectome Data Public—HCP Wiki. (n.d.). Retrieved September 9, 2020, from <https://wiki.humanconnectome.org/display/PublicData/HCP+Data+Release+Updates%3A+Known+Issues+and+Planned+fixes>
- HCP Subjects with Identified Quality Control Issues (QC_Issue measure codes explained). (n.d.). Retrieved September 9, 2020, from <https://wiki.humanconnectome.org/pages/viewpage.action?pageId=88901591>
- Heckner, M. K., Cieslik, E. C., Eickhoff, S. B., Camilleri, J. A., Hoffstaedter, F., & Langner, R. (2020). *The Aging Brain and Executive Functions Revisited: Implications from Meta-Analytic and Functional-Connectivity Evidence* [Preprint]. Neuroscience. <https://doi.org/10.1101/2020.07.15.204941>
- Jaccard, P. (1901). *Étude comparative de la distribution florale dans une portion des Alpes et du Jura* [Text/html,application/pdf]. <https://doi.org/10.5169/SEALS-266450>

- Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2), 143–156.
[https://doi.org/10.1016/s1361-8415\(01\)00036-6](https://doi.org/10.1016/s1361-8415(01)00036-6)
- Jenkinson, Mark, Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2), 825–841. [https://doi.org/10.1016/s1053-8119\(02\)91132-8](https://doi.org/10.1016/s1053-8119(02)91132-8)
- Jovicich, J., Czanner, S., Greve, D., Haley, E., van der Kouwe, A., Gollub, R., Kennedy, D., Schmitt, F., Brown, G., MacFall, J., Fischl, B., & Dale, A. (2006). Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data. *NeuroImage*, 30(2), 436–443.
<https://doi.org/10.1016/j.neuroimage.2005.09.046>
- Kampa, M., Schick, A., Sebastian, A., Wessa, M., Tüscher, O., Kalisch, R., & Yuen, K. (2020). Replication of fMRI group activations in the neuroimaging battery for the Mainz Resilience Project (MARP). *NeuroImage*, 204, 116223.
<https://doi.org/10.1016/j.neuroimage.2019.116223>
- Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, 55(4), 352–358.
<https://doi.org/10.1037/h0043688>
- Lahr, J., Minkova, L., Tabrizi, S. J., Stout, J. C., Klöppel, S., Scheller, E., & the TrackOn-HD Investigators. (2018). Working Memory-Related Effective Connectivity in Huntington's Disease Patients. *Frontiers in Neurology*, 9, 370.
<https://doi.org/10.3389/fneur.2018.00370>
- Laird, A. R., Fox, P. M., Price, C. J., Glahn, D. C., Uecker, A. M., Lancaster, J. L., Turkeltaub, P. E., Kochunov, P., & Fox, P. T. (2005). ALE meta-analysis: Controlling the false discovery rate and performing statistical contrasts. *Human Brain Mapping*, 25(1), 155–164. <https://doi.org/10.1002/hbm.20136>
- Laird, A. R., McMillan, K. M., Lancaster, J. L., Kochunov, P., Turkeltaub, P. E., Pardo, J. V., & Fox, P. T. (2005). A comparison of label-based review and ALE meta-analysis in the Stroop task. *Human Brain Mapping*, 25(1), 6–21.
<https://doi.org/10.1002/hbm.20129>
- Langner, R., & Eickhoff, S. B. (2013). Sustaining attention to simple tasks: A meta-analytic review of the neural mechanisms of vigilant attention. *Psychological Bulletin*, 139(4), 870–900. <https://doi.org/10.1037/a0030694>
- Langner, R., Leiberg, S., Hoffstaedter, F., & Eickhoff, S. B. (2018). Towards a human self-regulation system: Common and distinct neural signatures of emotional and behavioural control. *Neuroscience & Biobehavioral Reviews*, 90, 400–410.
<https://doi.org/10.1016/j.neubiorev.2018.04.022>
- 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>
- Mencarelli, L., Francesco, N., Davide, M., Arianna, M., Simone, R., Alessandro, R., & Emiliano, S. (2019). Stimuli, presentation modality, and load-specific brain

- activity patterns during n-back task. *Human Brain Mapping*, hbm.24633.
<https://doi.org/10.1002/hbm.24633>
- Migo, E. M., Mitterschiffthaler, M., O'Daly, O., Dawson, G. R., Dourish, C. T., Craig, K. J., Simmons, A., Wilcock, G. K., McCulloch, E., Jackson, S. H. D., Kopelman, M. D., Williams, S. C. R., & Morris, R. G. (2015). Alterations in working memory networks in amnesic mild cognitive impairment. *Aging, Neuropsychology, and Cognition*, 22(1), 106–127. <https://doi.org/10.1080/13825585.2014.894958>
- Miller, K. M., Price, C. C., Okun, M. S., Montijo, H., & Bowers, D. (2009). Is the N-Back Task a Valid Neuropsychological Measure for Assessing Working Memory? *Archives of Clinical Neuropsychology*, 24(7), 711–717.
<https://doi.org/10.1093/arclin/acp063>
- Minzenberg, M. J., Laird, A. R., Thelen, S., Carter, C. S., & Glahn, D. C. (2009). Meta-analysis of 41 Functional Neuroimaging Studies of Executive Function in Schizophrenia. *Archives of General Psychiatry*, 66(8), 811.
<https://doi.org/10.1001/archgenpsychiatry.2009.91>
- Morelli, S. A., Sacchet, M. D., & Zaki, J. (2015). Common and distinct neural correlates of personal and vicarious reward: A quantitative meta-analysis. *NeuroImage*, 112, 244–253. <https://doi.org/10.1016/j.neuroimage.2014.12.056>
- Müller, V. I., Cieslik, E. C., Laird, A. R., Fox, P. T., Radua, J., Mataix-Cols, D., Tench, C. R., Yarkoni, T., Nichols, T. E., Turkeltaub, P. E., Wager, T. D., & Eickhoff, S. B. (2018). Ten simple rules for neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews*, 84, 151–161. <https://doi.org/10.1016/j.neubiorev.2017.11.012>
- Müller, V. I., Cieslik, E. C., Serbanescu, I., Laird, A. R., Fox, P. T., & Eickhoff, S. B. (2017). Altered Brain Activity in Unipolar Depression Revisited: Meta-analyses of Neuroimaging Studies. *JAMA Psychiatry*, 74(1), 47.
<https://doi.org/10.1001/jamapsychiatry.2016.2783>
- Nee, D. E. (2019). fMRI replicability depends upon sufficient individual-level data. *Communications Biology*, 2(1), 130. <https://doi.org/10.1038/s42003-019-0378-6>
- Onwuegbuzie, A. J. (2000). *Running head: FRAMEWORK FOR INTERNAL AND EXTERNAL VALIDITY*. 63.
- Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping*, 25(1), 46–59. <https://doi.org/10.1002/hbm.20131>
- Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E., & Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18(2), 115–126. <https://doi.org/10.1038/nrn.2016.167>
- Poudel, R., Riedel, M. C., Salo, T., Flannery, J. S., Hill-Bowen, L. D., Eickhoff, S. B., Laird, A. R., & Sutherland, M. T. (2020). Common and distinct brain activity associated with risky and ambiguous decision-making. *Drug and Alcohol Dependence*, 209, 107884. <https://doi.org/10.1016/j.drugalcdep.2020.107884>
- Price, C. J., Devlin, J. T., Moore, C. J., Morton, C., & Laird, A. R. (2005). Meta-analyses of object naming: Effect of baseline. *Human Brain Mapping*, 25(1), 70–82.
<https://doi.org/10.1002/hbm.20132>

- Price, C. J., Moore, C. J., & Friston, K. J. (1997). Subtractions, conjunctions, and interactions in experimental design of activation studies. *Human Brain Mapping*, 5(4), 264–272. [https://doi.org/10.1002/\(SICI\)1097-0193\(1997\)5:4<264::AID-HBM11>3.0.CO;2-E](https://doi.org/10.1002/(SICI)1097-0193(1997)5:4<264::AID-HBM11>3.0.CO;2-E)
- Raichle, M. E. (1998). Imaging the mind. *Seminars in Nuclear Medicine*, 28(4), 278–289. [https://doi.org/10.1016/S0001-2998\(98\)80033-0](https://doi.org/10.1016/S0001-2998(98)80033-0)
- Richter, M., Amunts, K., Mohlberg, H., Bludau, S., Eickhoff, S. B., Zilles, K., & Caspers, S. (2019). Cytoarchitectonic segregation of human posterior intraparietal and adjacent parieto-occipital sulcus and its relation to visuomotor and cognitive functions. *Cerebral Cortex*, 29(3), 1305–1327. <https://doi.org/10.1093/cercor/bhy245>
- Rodríguez-Cano, E., Alonso-Lana, S., Sarró, S., Fernández-Corcuera, P., Goikolea, J. M., Vieta, E., Maristany, T., Salvador, R., McKenna, P. J., & Pomarol-Clotet, E. (2017). Differential failure to deactivate the default mode network in unipolar and bipolar depression. *Bipolar Disorders*, 19(5), 386–395. <https://doi.org/10.1111/bdi.12517>
- Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A. R., Schulz, J. B., Fox, P. T., & Eickhoff, S. B. (2012). Modelling neural correlates of working memory: A coordinate-based meta-analysis. *NeuroImage*, 60(1), 830–846. <https://doi.org/10.1016/j.neuroimage.2011.11.050>
- Ruan, J., Bludau, S., Palomero-Gallagher, N., Caspers, S., Mohlberg, H., Eickhoff, S. B., Seitz, R. J., & Amunts, K. (2018). Cytoarchitecture, probability maps, and functions of the human supplementary and pre-supplementary motor areas. *Brain Structure and Function*, 223(9), 4169–4186. <https://doi.org/10.1007/s00429-018-1738-6>
- Salimi-Khorshidi, G., Smith, S. M., Keltner, J. R., Wager, T. D., & Nichols, T. E. (2009). Meta-analysis of neuroimaging data: A comparison of image-based and coordinate-based pooling of studies. *NeuroImage*, 45(3), 810–823. <https://doi.org/10.1016/j.neuroimage.2008.12.039>
- Salo, T., Yarkoni, T., Nichols, T. E., Kent, J. D., Gorgolewski, K. J., Glerean, E., Bottenhorn, K. L., Bilgel, M., Wright, J., Reeders, P., Nielson, D. N., Yanes, J. A., Pérez, A., & Laird, A. R. (2020). *neurostuff/NiMARE: 0.0.3*. Zenodo. <https://doi.org/10.5281/zenodo.3892078>
- Sapara, A., ffytche, D. H., Birchwood, M., Cooke, M. A., Fannon, D., Williams, S. C. R., Kuipers, E., & Kumari, V. (2014). Preservation and compensation: The functional neuroanatomy of insight and working memory in schizophrenia. *Schizophrenia Research*, 152(1), 201–209. <https://doi.org/10.1016/j.schres.2013.11.026>
- Scheperjans, F., Eickhoff, S. B., Homke, L., Mohlberg, H., Hermann, K., Amunts, K., & Zilles, K. (2008). Probabilistic Maps, Morphometry, and Variability of Cytoarchitectonic Areas in the Human Superior Parietal Cortex. *Cerebral Cortex*, 18(9), 2141–2157. <https://doi.org/10.1093/cercor/bhm241>
- Scheperjans, F., Hermann, K., Eickhoff, S. B., Amunts, K., Schleicher, A., & Zilles, K. (2008). Observer-Independent Cytoarchitectonic Mapping of the Human Superior Parietal Cortex. *Cerebral Cortex*, 18(4), 846–867. <https://doi.org/10.1093/cercor/bhm116>

- Schlösser, R. G. M., Nenadic, I., Wagner, G., Güllmar, D., von Consbruch, K., Köhler, S., Schultz, C. C., Koch, K., Fitzek, C., Matthews, P. M., Reichenbach, J. R., & Sauer, H. (2007). White matter abnormalities and brain activation in schizophrenia: A combined DTI and fMRI study. *Schizophrenia Research*, 89(1–3), 1–11. <https://doi.org/10.1016/j.schres.2006.09.007>
- Shaffer, J. P. (1995). Multiple Hypothesis Testing. *Annual Review of Psychology*, 46(1), 561–584. <https://doi.org/10.1146/annurev.ps.46.020195.003021>
- Simmonds, D. J., Pekar, J. J., & Mostofsky, S. H. (2008). Meta-analysis of Go/No-go tasks demonstrating that fMRI activation associated with response inhibition is task-dependent. *Neuropsychologia*, 46(1), 224–232. <https://doi.org/10.1016/j.neuropsychologia.2007.07.015>
- Sørensen, T. J. (1948). *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons*. I kommission hos E. Munksgaard; /z-wcorg/.
- Stark, C. E. L., & Squire, L. R. (2001). When zero is not zero: The problem of ambiguous baseline conditions in fMRI. *Proceedings of the National Academy of Sciences*, 98(22), 12760–12766. <https://doi.org/10.1073/pnas.221462998>
- Turkeltaub, P. E., Eden, G. F., Jones, K. M., & Zeffiro, T. A. (2002). Meta-analysis of the functional neuroanatomy of single-word reading: Method and validation. *NeuroImage*, 16(3 Pt 1), 765–780. <https://doi.org/10.1006/nimg.2002.1131>
- Turkeltaub, P. E., Eickhoff, S. B., Laird, A. R., Fox, M., Wiener, M., & Fox, P. (2012). Minimizing within-experiment and within-group effects in activation likelihood estimation meta-analyses. *Human Brain Mapping*, 33(1), 1–13. <https://doi.org/10.1002/hbm.21186>
- Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology*, 1(1), 62. <https://doi.org/10.1038/s42003-018-0073-z>
- Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E. J., Yacoub, E., & Ugurbil, K. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage*, 80, 62–79. <https://doi.org/10.1016/j.neuroimage.2013.05.041>
- Woolrich, M. W., Behrens, T. E. J., Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2004). Multilevel linear modelling for FMRI group analysis using Bayesian inference. *NeuroImage*, 21(4), 1732–1747. <https://doi.org/10.1016/j.neuroimage.2003.12.023>
- Woolrich, M. W., Ripley, B. D., Brady, M., & Smith, S. M. (2001). Temporal Autocorrelation in Univariate Linear Modeling of FMRI Data. *NeuroImage*, 14(6), 1370–1386. <https://doi.org/10.1006/nimg.2001.0931>
- Yaple, Z. A., Stevens, W. D., & Arsalidou, M. (2019). Meta-analyses of the n-back working memory task: FMRI evidence of age-related changes in prefrontal cortex involvement across the adult lifespan. *NeuroImage*, 196, 16–31. <https://doi.org/10.1016/j.neuroimage.2019.03.074>

Supplementary data

Overview of literature dataset

Supplementary Table 1. Overview of experiments included in the different meta-analyses.

Study	Contrasts (i.e. experiments)	Subjects (female)	Age (\pm SD)	Smoothing (FWHM)	Correction	Modality	Stimuli	Used in contrast analysis
Aguilar-Ortiz et al. 2019	2-back > baseline 2-back > 1-back 1-back > baseline	67 (64)	32.5 \pm 9.68		cFWE p < 0.05	visual	letter	(2-back > rest/ bsl.) > (1-back > rest/ bsl.)
Aguirre et al. 2019*	2-back > 0-back	29 (14)	32.72	8	cFWE p < 0.05	visual	letter	
Alain et al. 2010	2-back > 1-back	12 (7)	26.33 \pm 2.9	6	FWE p < 0.05	auditory	sound	
Alain et al. 2018	2-back > 1-back	41 (25)	25.05	6		visual/ auditory		
Allen et al. 2006	2-back > 0-back	10 (2)		7.2	cFWE p < 0.01	visual	letter	
Alonso-Lana et al. 2016	2-back > baseline 2-back > 1-back 1-back > baseline	28 (16)	44.01 \pm 6.03	5	cFWE p < 0.05	visual	letter	(2-back > rest/ bsl.) > (1-back > rest/ bsl.)
Bleich-Cohen et al. 2014	2-back > 0-back	20 (8)	26.4 \pm 2.7		FDR p < 0.05	visual	numbers	
Boller et al. 2017	2-back > rest 2-back > 1-back 1-back > rest	32 (25)	68.59 \pm 6.5	9	FWE p < 0.05	visual	letter	(2-back > rest/ bsl.) > (1-back > rest/ bsl.)
Campanella et al. 2013*	2-back > 0-back	16 (9)	21.6 \pm 2.6		uncorrected p < 0.001	visual	numbers	
Cerasa et al. 2008	2-back > 0-back	30 (0)	30.79	6	cFWE p < 0.05	visual	shapes	
Choo et al. 2005	2-back > 1-back	12		8	uncorrected p < 0.001	visual	letter	

Ciesielski et al. 2006	2-back > baseline	10 (5)	23.5 ± 2.29	6	uncorrected p < 0.001	visual	figures	
Clark et al. 2017	2-back > 1-back	63 (35)	30.91 ± 6.01	7	cFWE p < 0.05	visual	shapes/ letter	
Daamen et al. 2015* c	2-back > 0-back 2-back > rest 0-back > rest	73 (28)	26.51 ± 0.53	10	FWE p < 0.05	visual	letter	
Daamen et al. 2015* v	2-back > 0-back 2-back > rest 0-back > rest	73 (29)	26.5 ± 0.49	10	FWE p < 0.05	visual	letter	
Deckersbach et al. 2008	2-back > rest	17 (17)	25.6 ± 5.9	6	uncorrected p < 0.001	visual	letter	(2-back > rest/ bsl.) > (1-back > rest/ bsl.)
Dima et al. 2014	2-back > 0-back 1-back > 0-back	40 (20)	31.5 ± 10.4	8	cFWE p < 0.05	visual	letter	(2-back > 0-back) > (1-back > 0-back)
Döhnelt et al. 2008	2-back > rest	16 (8)	61 ± 10.2	8	uncorrected p < 0.001	visual	figures	(2-back > rest/ bsl.) > (1-back > rest/ bsl.)
Dores et al. 2017	2-back > rest	10 (4)	27.1 ± 2.89		FDR p < 0.05	visual	shapes	(2-back > rest/ bsl.) > (1-back > rest/ bsl.)
Drapier et al. 2008	2-back > 0-back 1-back > 0-back	20 (10)	41.9 ± 11.6			visual	letter	(2-back > 0-back) > (1-back > 0-back)
Duggirala et al. 2016 A	2-back > 0-back	50 (22)	23.62 ± 3.17	4		visual	words	
Esteves et al. 2018	2-back > 0-back 1-back > 0-back	31 (10)	60.29 ± 7.71	8	FWE p < 0.05	visual	letter	(2-back > 0-back) > (1-back > 0-back)
Fernández-Corcuera et al. 2013	2-back > baseline	41 (17)	40.27 ± 9.8		cFWE p < 0.05	visual	letter	
Forn et al. 2007	2-back > 0-back	10 (5)	31.1	8	uncorrected p < 0.001	auditory	letter	(2-back > 0-back) > (1-back > 0-back)
Fuentes-Claramonte et al. 2019	2-back > rest	36 (16)	41.19 ± 11.99	5	cFWE p < 0.05	visual	letter	
Fukuda et al. 2019* h	2-back > 0-back 2-back > baseline 0-back > baseline	24 (8)	23.54 ± 5.35	6	uncorrected p < 0.001	visual	numbers	

Fukuda et al. 2019*	2-back > 0-back 2-back > baseline 0-back > baseline	24 (8)	25.29 ± 4.77	6	uncorrected p < 0.001	visual	numbers	
Garrett et al. 2011	2-back > 0-back 1-back > 0-back	19 (6)	34.85 ± 12.54	4	cFWE p < 0.01	visual	letter	(2-back > 0-back) > (1-back > 0-back)
Gillis et al. 2016	2-back > 0-back	15 (0)	25.13 ± 4.55	4	cFWE	visual	letter/ visuospatial	
Goikolea et al. 2019	2-back > baseline	31 (15)	31.06 ± 8.76	8	FWE p < 0.05	visual	letter	
Gropman et al. 2013	2-back > 1-back	21 (14)	31.8 ± 2.7	8	FWE p < 0.01	visual	letter	
Habel et al. 2007	2-back > 0-back	21 (0)	30.77 ± 9.65	10	FDR p < 0.05	visual	letter	
Habel et al. 2007 - Koch et al. 2007*	0-back > rest	47	31.4 ± 10.4	10		visual	letter	
Harding et al. 2016*	2-back > 0-back 2-back > baseline 0-back > baseline	34 (17)	33.6	5	cFWE p < 0.05	visual	letter	
Heinzel et al. 2016	2-back > rest 0-back > rest 1-back > rest	29 (18)	66.04	8	FWE p < 0.05	visual	numbers	(2-back > rest/ bsl.) > (1-back > rest/ bsl.)
Honey et al. 2000	2-back > 0-back	20 (0)	39.3 ± 13.6	7	uncorrected p < 0.0005	visual	letter	
Honey et al. 2003	2-back > 0-back	27 (6)	35.1 ± 9.9		uncorrected p < 0.001	visual	letter	
Huang et al. 2016	2-back > 1-back 1-back > 0-back	18 (12)	43.17 ± 6.48	6		visual	shapes	
Jiang et al. 2015 o	2-back > 0-back	20 (10)	51.8 ± 5.9	8	FWE p < 0.05	visual	numbers	
Jiang et al. 2015 y	2-back > 0-back	20	23.1 ± 3.1	8	FWE p < 0.05	visual	numbers	
Johannsen et al. 2013	2-back > baseline	12 (8)	26.1 ± 4.7	8		visual	letter	(2-back > rest/ bsl.) > (1-back > rest/ bsl.)
Jung et al. 2018*	2-back > 0-back	24 (11)	22.9 ± 2.7	6	FWE p < 0.05	visual	numbers	
Kaminski et al. 2020*	2-back > 0-back	41 (12)	34.39 ± 8.53	6	FWE p < 0.05	visual	numbers	

	2-back > baseline							
	0-back > baseline							
Kim et al. 2006	2-back > rest	12 (3)	34.4 ± 9.5		cFWE p < 0.005	visual	letter	(2-back > rest/ bsl.) > (1-back > rest/ bsl.)
King et al. 2015*	2-back > 0-back	17 (11)	23.24 ± 5.89	5	cFWE p < 0.05	visual	letter	
	0-back > baseline							
Knops et al. 2006	2-back > 1-back	13 (0)	27 ± 7.65	12	FDR p < 0.05	visual	letter/ numbers	
Koppelstaetter et al. 2008	2-back > 0-back	15 (0)		8	cFWE p < 0.05	visual	letter	
Korsnes et al. 2013	2-back > 1-back	11 (11)	30.2 ± 5.95	8	FDR p < 0.05	visual	numbers	
Krug et al. 2008*	0-back > rest	85 (27)	23.27	6		visual	letter	
Kumari et al. 2006	2-back > 0-back	13 (0)	33.31 ± 6.85	10	FWE p < 0.05	visual	shapes	(2-back > 0-back) > (1-back > 0-back)
	0-back > rest							
	1-back > 0-back							
Lahr et al. 2018*	2-back > 0-back	83 (48)	49.11 ± 10.33			visual	letter	(2-back > 0-back) > (1-back > 0-back)
	1-back > 0-back							
	2-back > 1-back							
Lamp et al. 2016	1-back > baseline	16 (9)	23.94 ± 2.49	5	FWE p < 0.05	visual	shapes	
Lee et al. 2013	1-back > rest	14 (5)	64.8 ± 4.2	8	FWE p < 0.025	visual	numbers	
Leung & Alain 2011 A	2-back > 1-back	16 (9)	25.19 ± 5.13	6	FWE p < 0.05	auditory	sound	
L. Li et al. 2014	2-back > rest	15 (15)	19.56 ± 0.81	8	uncorrected p < 0.001	visual	letter	(2-back > rest/ bsl.) > (1-back > rest/ bsl.)
	0-back > rest							
	1-back > rest							
L. Li et al. 2019a*	2-back > 0-back	24 (24)	25.63 ± 0.65	8	uncorrected p < 0.001	visual	letter	
	2-back > rest							(2-back > rest/ bsl.) > (1-back > rest/ bsl.)
	0-back > rest							
	2-back > 1-back							
	1-back > rest							

X. Li et al. 2019	2-back > 0-back	24 (8)	24.08 ± 4.34	8	cFWE p < 0.05	visual	shapes/ letter	
Lim et al. 2008	1-back > rest	12 (7)	68.6 ± 6.2	8	FDR p < 0.01	visual	letter	
Luo et al. 2014	2-back > 0-back	25 (0)	23.14 ± 1.83	8	FWE p < 0.05	visual	faces	
Malisza et al. 2005	1-back > 0-back	6		8		visual	shapes	
Marquand et al. 2008	2-back > 0-back	20 (13)	43.7 ± 8.3	8		visual	letter	(2-back > 0-back) > (1-back > 0-back)
Matsuo et al. 2007	2-back > 0-back 1-back > 0-back	15 (9)	37.7 ± 12.1	5	cFWE p < 0.05	visual	shapes numbers	(2-back > 0-back) > (1-back > 0-back)
McAllister et al. 1999	2-back > 1-back 1-back > 0-back	11 (7)	30.6 ± 11.2	15	uncorrected p < 0.001	auditory	letter	
McGeown et al. 2008	1-back > 0-back	9 (6)	75.11 ± 1.62	8	FWE p < 0.05	visual	words	
Meisenzahl et al. 2006	2-back > 0-back	12 (1)	33.58 ± 9.27	8	uncorrected p < 0.001	visual	letter	(2-back > 0-back) > (1-back > 0-back)
Migo et al. 2014	2-back > 0-back 1-back > 0-back	11 (4)	70.27 ± 6.2	8	cFWE p < 0.05	visual	letter	(2-back > 0-back) > (1-back > 0-back)
Miró-Padilla et al. 2019*	2-back > 0-back	52 (31)	22.6 ± 1.45	8	uncorrected p < 0.001	visual	letter	
Monks et al. 2004	2-back > 0-back	12 (0)	45.6 ± 3.52			visual	letter	
Nebel et al. 2005 f	1-back > baseline	19 (7)	30.3 ± 3.2	9	FWE p < 0.05	visual	letter/ pictures	
Nebel et al. 2005 s	1-back > baseline	17 (11)	26.94 ± 5.5	9	FWE p < 0.05	visual	letter/ pictures	
Ogg et al. 2008	0-back > rest	30 (17)	24.2	6	FWE p < 0.05	visual	letter	
Park et al. 2016	2-back > 0-back	45 (23)	22.87	8	FWE p < 0.01	visual	shapes	
Pfefferbaum et al. 2001	2-back > 0-back 2-back > rest 0-back > rest	10 (0)	60.2 ± 12.8	5	FWE p < 0.05	visual	letter	(2-back > 0-back) > (1-back > 0-back) (2-back > rest/ bsl.) > (1-back > rest/ bsl.)
Philip et al. 2016	2-back > rest 0-back > rest	13 (9)	30 ± 9	6	FWE p < 0.05	visual	letter	(2-back > rest/ bsl.) > (1-back > rest/ bsl.)
Pomarol-Clotet et al. 2008	2-back > baseline	32 (11)	41.03 ± 11.04		cFWE p < 0.05	visual	letter	

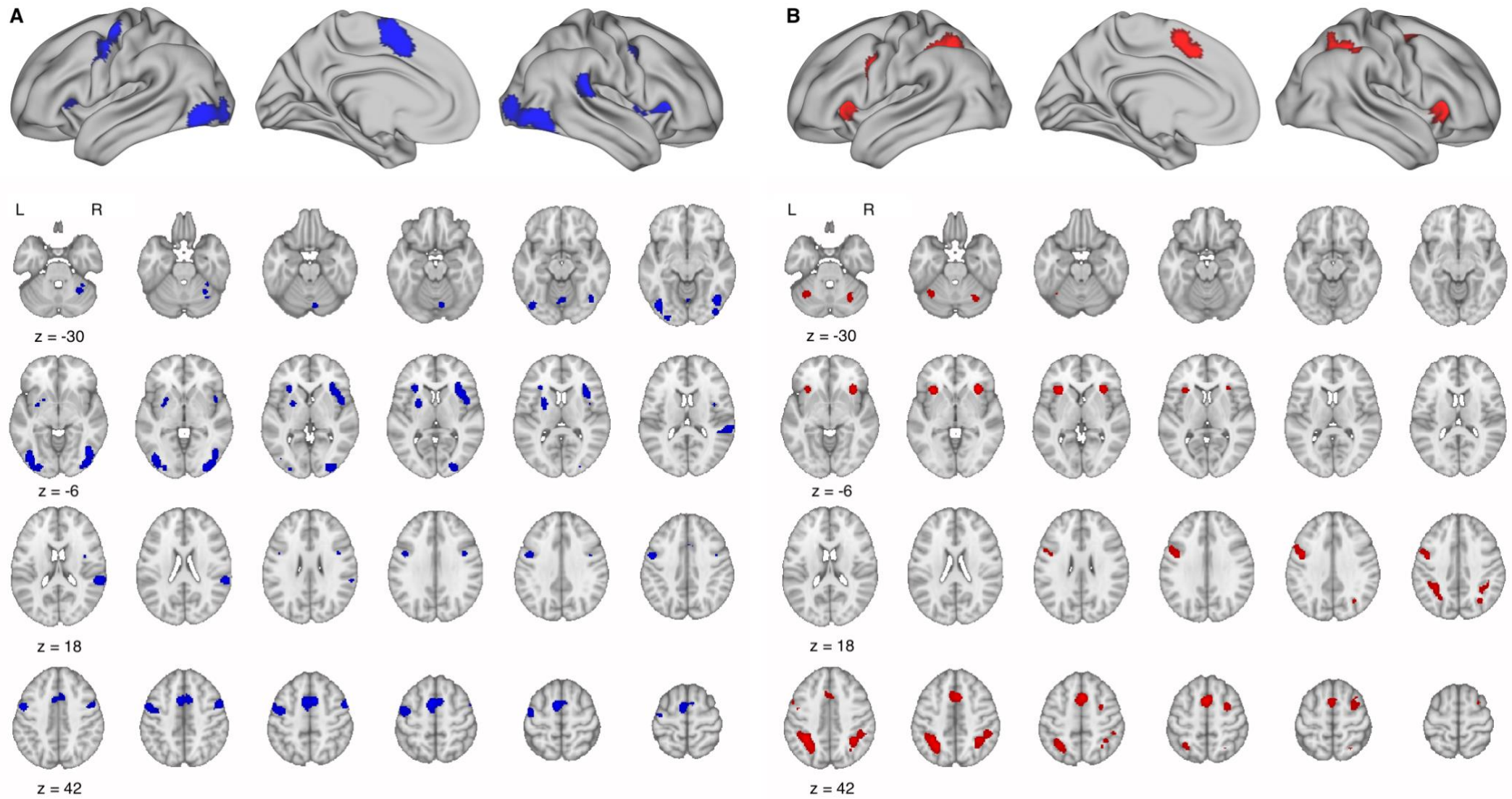
Qin et al. 2009	2-back > 0-back	27 (27)	20.52	8	cFWE p < 0.05	visual	numbers	
Ragland et al. 2002 A	2-back > 0-back 2-back > 1-back 1-back > 0-back	11 (5)	32.2	12	FWE p < 0.05	visual	letter	(2-back > 0-back) > (1-back > 0-back)
Rama et al. 2001	2-back > 0-back 1-back > 0-back	8 (8)	22			auditory	words	(2-back > 0-back) > (1-back > 0-back)
Ricciardi et al. 2006 A	1-back > rest	6 (6)	28 ± 1	3.4	FDR	visual	shapes	
Richter et al. 2013	2-back > 0-back	34 (17)	23.8	8	FWE p < 0.05	visual	faces	
Rodriguez-Cano et al. 2014	2-back > baseline	52 (32)	46.25 ± 10.21			visual	letter	
Rodriguez-Cano et al. 2017	2-back > baseline 1-back > baseline	26	46.77 ± 11.18	5	cFWE p < 0.05	visual	letter	(2-back > rest/ bsl.) > (1-back > rest/ bsl.)
Salavert et al. 2018	2-back > baseline 1-back > baseline	41 (13)	31.7 ± 9.6	5		visual	letter	(2-back > rest/ bsl.) > (1-back > rest/ bsl.)
Sánchez-Carrión et al. 2008	2-back > 0-back	14	24.2 ± 4.7	10	FDR p < 0.001	visual	numbers	
Sapara et al. 2014	2-back > 0-back 2-back > rest 0-back > rest 1-back > rest 1-back > 0-back	20 (5)	31.95 ± 7.6	8	cFWE p < 0.05	visual	shapes	(2-back > 0-back) > (1-back > 0-back) (2-back > rest/ bsl.) > (1-back > rest/ bsl.)
Scheller et al. 2017	2-back > 0-back 1-back > 0-back	34	68.82 ± 5.33	6	FWE p < 0.05	visual	letter	(2-back > 0-back) > (1-back > 0-back)
Scheuerecker et al. 2008	2-back > 0-back	23 (4)	32.6 ± 9.9	8	cFWE p < 0.05	visual	letter	
Schlagenhauf et al. 2008*	0-back > rest	10 (2)	33.8 ± 12.5	8		visual	numbers	
Schmidt et al. 2015	2-back > 0-back	32	24.6	8	FWE p < 0.05	visual	letter	
Schneider et al. 2007*	0-back > rest	81	30.9 ± 8.3	10		visual	letter	
Schneiders et al. 2011	2-back > 0-back	48 (26)	23.67	6	FDR p < 0.01	visual	shapes	

Seo et al. 2012	2-back > 0-back	22 (22)	38.27 ± 8.48		FDR p < 0.01	visual	letter	
Seo et al. 2014	2-back > 0-back	34 (34)	59.3 ± 5.2	8	FWE p < 0.05	visual	letter	
Shen et al. 1999	1-back > rest	9 (3)				visual	shapes	
Smits et al. 2009*	0-back > rest	12 (4)	27.8 ± 10			auditory	numbers	
Spreng et al. 2014	2-back > rest	36 (19)	22.3 ± 3.8	6		visual	faces	
Thomas et al. 2005	2-back > baseline	16	37.6 ± 6.3	6	FWE p < 0.05	visual	letter	(2-back > rest/ bsl.) > (1-back > rest/ bsl.)
Thornton and Conway 2013	2-back > 1-back	14 (9)	22 ± 2.45	6	cFWE p < 0.05	visual	visual	
Vacchi et al. 2017*	2-back > 0-back 2-back > 1-back 1-back > 0-back	24 (12)	37.6 ± 12.2		cFWE p < 0.05	visual	letter	(2-back > 0-back) > (1-back > 0-back)
van der Horn et al. 2016*	2-back > 0-back 2-back > 1-back 1-back > 0-back	20 (7)	34	8	uncorrected p < 0.001	visual	letter	(2-back > 0-back) > (1-back > 0-back)
Waiter et al. 2009	2-back > 0-back	37 (17)	69.8 ± 0.4	6	FWE p < 0.05	visual	letter	
Walitt et al. 2016	2-back > 0-back	13 (13)	44.2 ± 11.2	8	cFWE p < 0.05	visual	letter	
Wesley et al. 2017	2-back > 1-back 1-back > 0-back	11 (7)	28.8 ± 7.8	8	uncorrected p < 0.001	visual	letter	
Wishart et al. 2006	2-back > 0-back	22 (11)	68.5 ± 13.3	10	cFWE p < 0.05	auditory	letter	
Wu et al. 2017*	2-back > 0-back 2-back > rest 0-back > rest	45 (21)	24.07 ± 4.83	6	FWE p < 0.05	visual	numbers	
Yan et al. 2011 h	2-back > 0-back	28 (16)	20.4 ± 1.4	6	cFWE p < 0.001	visual	shapes	
Yan et al. 2011 s	2-back > 0-back	28 (16)	20.9 ± 1.5	6	cFWE p < 0.001	visual	shapes	
Yang et al. 2018	2-back > 0-back	24 (12)	22.1 ± 2.2	6	FWE p < 0.05	visual	letter	
Yoo et al. 2004 A	1-back > rest	14 (5)	26.3	6	FWE p < 0.05	visual	letter	
Yoo et al. 2005	2-back > 0-back	10 (2)	22.6 ± 1.4	6	FWE p < 0.05	visual	faces	

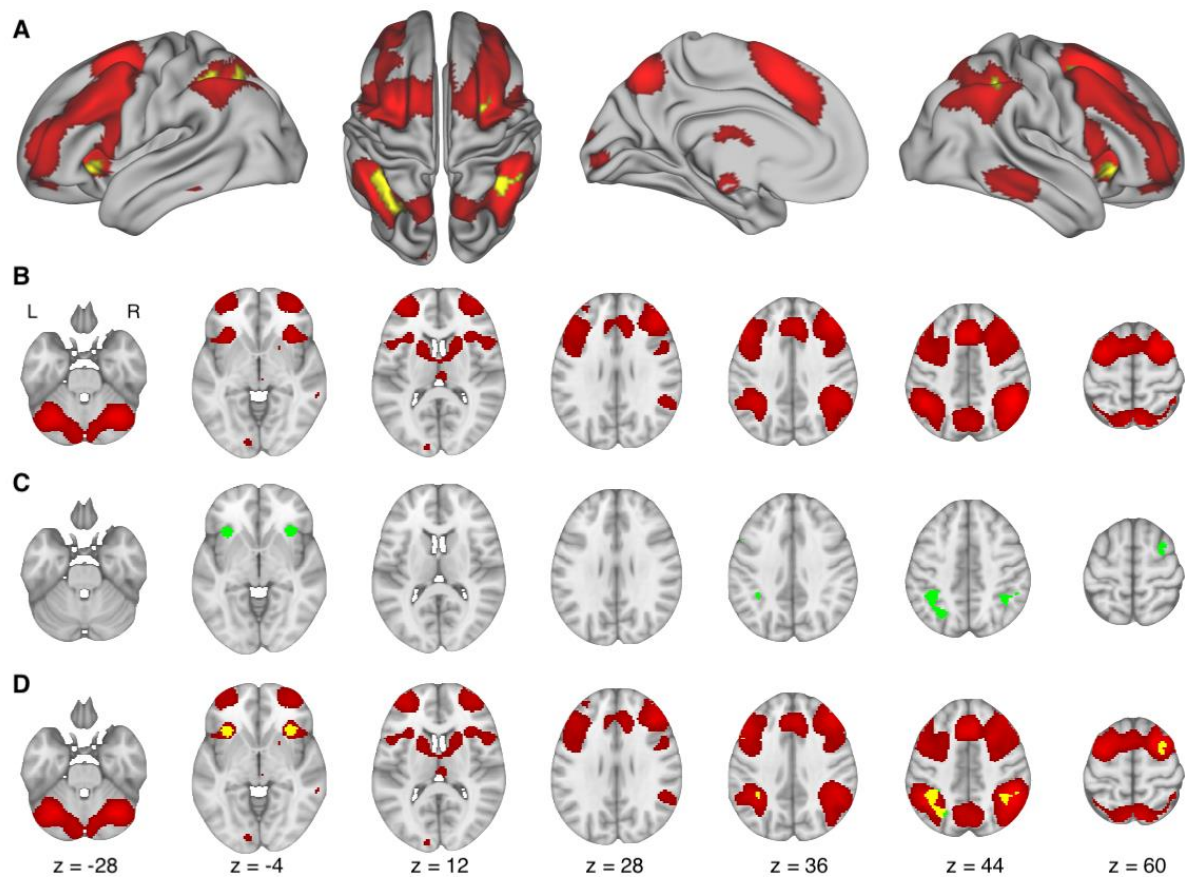
Ziemus et al. 2007	2-back > 0-back	9 (4)	44.2 ± 9.6	8	FWE p < 0.05	visual	letter	(2-back > 0-back) > (1-back > 0-back)
--------------------	-----------------	-------	------------	---	--------------	--------	--------	---------------------------------------

Note: Study names with an asterisk * indicate, that the authors kindly provided additional contrast coordinates, not reported in the original publication. Small letters after the author indicate different subject groups in publication. Large letter (A), indicate the “first” reported coordinates were used (in case of multiple experiments with the same group).

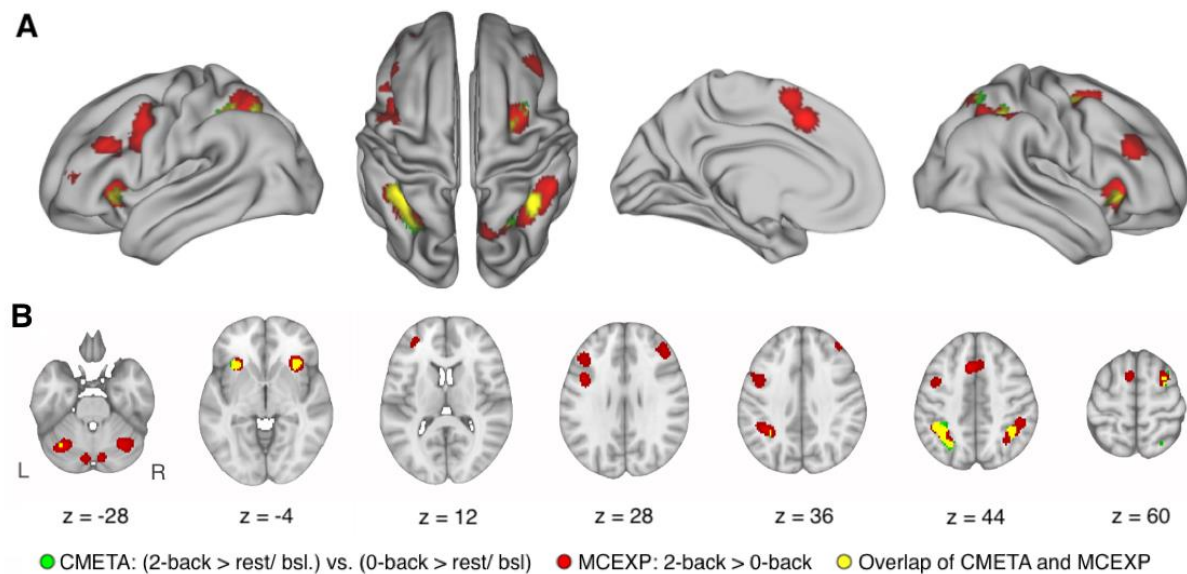
Additional literature-based meta-analyses results



Supplementary Figure 1. A) Meta-analysis across 0-back > rest/ baseline experiments (21); B) Meta-analysis across 2-back > rest/ baseline experiments (31); Axial slices in MNI space.



Supplementary Figure 2. Brain regions revealed by contrast between meta-analyses (2-back > rest/baseline) > (0-back > rest/baseline) (Cmeta) and by large sample 2-back > 0-back contrast. (A) and (D) show large-sample contrast in red; Cmeta in green; Overlap of MCexp and large-sample contrast in yellow. (B) large-sample contrast. (C) Cmeta. Top row, cortex maps. Axial slices in MNI space.



Supplementary Figure 3. Reduced 2-back > 0-back: Contrast between two meta-analyses (Cmeta) vs. meta-analysis across contrasts (MCexp). Cmeta: (2-back > rest/baseline) (21 experiments, matched to 0-back > rest/baseline experiments) vs. (0-back > rest/baseline) (21 experiments); MCexp: meta-analysis across 2-back > 0-back (21 experiments, matched to Cmeta experiments); (A) cortex maps. (B) Axial slices in MNI space.

Supplementary Bibliography

- Aguilar-Ortiz, S., Salgado-Pineda, P., Vega, D., Pascual, J. C., Marco-Pallarés, J., Soler, J., Brunel, C., Martin-Blanco, A., Soto, A., Ribas, J., Maristany, T., Sarró, S., Rodríguez-Fornells, A., Salvador, R., McKenna, P. J., & Pomarol-Clotet, E. (2020). Evidence for default mode network dysfunction in borderline personality disorder. *Psychological Medicine*, 50(10), 1746–1754. <https://doi.org/10.1017/S0033291719001880>
- Aguirre, N., Cruz-Gómez, Á. J., Miró-Padilla, A., Bueichekú, E., Broseta Torres, R., Ávila, C., Sanchis-Segura, C., & Forn, C. (2019). Repeated Working Memory Training Improves Task Performance and Neural Efficiency in Multiple Sclerosis Patients and Healthy Controls. *Multiple Sclerosis International*, 2019, 1–13. <https://doi.org/10.1155/2019/2657902>
- Alain, C., Khatamian, Y., He, Y., Lee, Y., Moreno, S., Leung, A. W. S., & Bialystok, E. (2018). Different neural activities support auditory working memory in musicians and bilinguals: Neural resources in musicians and bilinguals. *Annals of the New York Academy of Sciences*, 1423(1), 435–446. <https://doi.org/10.1111/nyas.13717>
- Alain, C., Shen, D., Yu, H., & Grady, C. (2010). Dissociable Memory- and Response-Related Activity in Parietal Cortex During Auditory Spatial Working Memory. *Frontiers in Psychology*, 1. <https://doi.org/10.3389/fpsyg.2010.00202>
- Allen, P. P., Cleare, A. J., Lee, F., Fusar-Poli, P., Tunstall, N., Fu, C. H. Y., Brammer, M. J., & McGuire, P. K. (2006). Effect of acute tryptophan depletion on pre-frontal engagement. *Psychopharmacology*, 187(4), 486–497. <https://doi.org/10.1007/s00213-006-0444-x>
- Alonso-Lana, S., Goikolea, J. M., Bonnin, C. M., Sarró, S., Segura, B., Amann, B. L., Monté, G. C., Moro, N., Fernandez-Corcuera, P., Maristany, T., Salvador, R., Vieta, E., Pomarol-Clotet, E., & McKenna, P. J. (2016). Structural and Functional Brain Correlates of Cognitive Impairment in Euthymic Patients with Bipolar Disorder. *PLOS ONE*, 11(7), e0158867. <https://doi.org/10.1371/journal.pone.0158867>
- Bleich-Cohen, M., Hendler, T., Weizman, R., Faragian, S., Weizman, A., & Poyurovsky, M. (2014). Working memory dysfunction in schizophrenia patients with obsessive-compulsive symptoms: An fMRI study. *European Psychiatry*, 29(3), 160–166. <https://doi.org/10.1016/j.eurpsy.2013.02.004>
- Boller, B., Mellah, S., Ducharme-Laliberté, G., & Belleville, S. (2017). Relationships between years of education, regional grey matter volumes, and working memory-related brain activity in healthy older adults. *Brain Imaging and Behavior*, 11(2), 304–317. <https://doi.org/10.1007/s11682-016-9621-7>
- Campanella, S., Peigneux, P., Petit, G., Lallemand, F., Saeremans, M., Noël, X., Metens, T., Nouali, M., De Tiège, X., De Witte, P., Ward, R., & Verbanck, P. (2013). Increased Cortical Activity in Binge Drinkers during Working Memory Task: A Preliminary Assessment through a Functional Magnetic Resonance Imaging Study. *PLoS ONE*, 8(4), e62260. <https://doi.org/10.1371/journal.pone.0062260>
- Cerasa, A., Gioia, M. C., Fera, F., Passamonti, L., Liguori, M., Lanza, P., Muglia, M., Magariello, A., & Quattrone, A. (2008). Ventro-lateral prefrontal activity during

- working memory is modulated by MAO A genetic variation. *Brain Research*, 1201, 114–121. <https://doi.org/10.1016/j.brainres.2008.01.048>
- Choo, W.-C., Lee, W.-W., Venkatraman, V., Sheu, F.-S., & Chee, M. W. L. (2005). Dissociation of cortical regions modulated by both working memory load and sleep deprivation and by sleep deprivation alone. *NeuroImage*, 25(2), 579–587. <https://doi.org/10.1016/j.neuroimage.2004.11.029>
- Ciesielski, K. T., Lesnik, P. G., Savoy, R. L., Grant, E. P., & Ahlfors, S. P. (2006). Developmental neural networks in children performing a Categorical N-Back Task. *NeuroImage*, 33(3), 980–990. <https://doi.org/10.1016/j.neuroimage.2006.07.028>
- Clark, C. M., Lawlor-Savage, L., & Goghari, V. M. (2017). Comparing brain activations associated with working memory and fluid intelligence. *Intelligence*, 63, 66–77. <https://doi.org/10.1016/j.intell.2017.06.001>
- Daamen, M., Bäuml, J. G., Scheef, L., Sorg, C., Busch, B., Baumann, N., Bartmann, P., Wolke, D., Wohlschläger, A., & Boecker, H. (2015). Working memory in preterm-born adults: Load-dependent compensatory activity of the posterior default mode network: Working Memory in Preterm-Born Adults. *Human Brain Mapping*, 36(3), 1121–1137. <https://doi.org/10.1002/hbm.22691>
- Deckersbach, T., Rauch, S. L., Buhlmann, U., Ostacher, M. J., Beucke, J.-C., Nierenberg, A. A., Sachs, G., & Dougherty, D. D. (2008). An fMRI investigation of working memory and sadness in females with bipolar disorder: A brief report. *Bipolar Disorders*, 10(8), 928–942. <https://doi.org/10.1111/j.1399-5618.2008.00633.x>
- Dima, D., Jogia, J., & Frangou, S. (2014). Dynamic causal modeling of load-dependent modulation of effective connectivity within the verbal working memory network: Brain Connectivity in Increasing Memory Load. *Human Brain Mapping*, 35(7), 3025–3035. <https://doi.org/10.1002/hbm.22382>
- Döhnell, K., Sommer, M., Ibach, B., Rothmayr, C., Meinhardt, J., & Hajak, G. (2008). Neural correlates of emotional working memory in patients with mild cognitive impairment. *Neuropsychologia*, 46(1), 37–48. <https://doi.org/10.1016/j.neuropsychologia.2007.08.012>
- Dores, A. R., Barbosa, F., Carvalho, I. P., Almeida, I., Guerreiro, S., da Rocha, B. M., de Sousa, L., & Castro-Caldas, A. (2017). Study of behavioural and neural bases of visuo-spatial working memory with an fMRI paradigm based on an n-back task. *Journal of Neuropsychology*, 11(1), 122–134. <https://doi.org/10.1111/jnp.12076>
- Drapier, D., Surguladze, S., Marshall, N., Schulze, K., Fern, A., Hall, M.-H., Walshe, M., Murray, R. M., & McDonald, C. (2008). Genetic Liability for Bipolar Disorder Is Characterized by Excess Frontal Activation in Response to a Working Memory Task. *Biological Psychiatry*, 64(6), 513–520. <https://doi.org/10.1016/j.biopsych.2008.04.038>
- Duggirala, S. X., Saharan, S., Raghunathan, P., & Mandal, P. K. (2016). Stimulus-dependent modulation of working memory for identity monitoring: A functional MRI study. *Brain and Cognition*, 102, 55–64. <https://doi.org/10.1016/j.bandc.2015.12.006>
- Esteves, M., Magalhães, R., Marques, P., Castanho, T. C., Portugal-Nunes, C., Soares, J. M., Almeida, A., Santos, N. C., Sousa, N., & Leite-Almeida, H. (2018). Functional

- Hemispheric (A)symmetries in the Aged Brain—Relevance for Working Memory. *Frontiers in Aging Neuroscience*, 10, 58.
<https://doi.org/10.3389/fnagi.2018.00058>
- Fernández-Corcuera, P., Salvador, R., Monté, G. C., Salvador Sarró, S., Goikolea, J. M., Amann, B., Moro, N., Sans-Sansa, B., Ortiz-Gil, J., Vieta, E., Maristany, T., McKenna, P. J., & Pomarol-Clotet, E. (2013). Bipolar depressed patients show both failure to activate and failure to de-activate during performance of a working memory task. *Journal of Affective Disorders*, 148(2–3), 170–178.
<https://doi.org/10.1016/j.jad.2012.04.009>
- Forn, C., Barros-Loscertales, A., Escudero, J., Benlloch, V., Campos, S., Antònia Parcet, M., & Ávila, C. (2007). Compensatory activations in patients with multiple sclerosis during preserved performance on the auditory N-back task. *Human Brain Mapping*, 28(5), 424–430. <https://doi.org/10.1002/hbm.20284>
- Fuentes-Claramonte, P., Martín-Subero, M., Salgado-Pineda, P., Alonso-Lana, S., Moreno-Alcázar, A., Argila-Plaza, I., Santo-Angles, A., Albajes-Eizagirre, A., Anguera-Camós, M., Capdevila, A., Sarró, S., McKenna, P. J., Pomarol-Clotet, E., & Salvador, R. (2019). Shared and differential default-mode related patterns of activity in an autobiographical, a self-referential and an attentional task. *PLOS ONE*, 14(1), e0209376. <https://doi.org/10.1371/journal.pone.0209376>
- Fukuda, Y., Katthagen, T., Deserno, L., Shayegan, L., Kaminski, J., Heinz, A., & Schlagenhauf, F. (2019). Reduced parietofrontal effective connectivity during a working-memory task in people with high delusional ideation. *Journal of Psychiatry & Neuroscience*, 44(3), 195–204. <https://doi.org/10.1503/jpn.180043>
- Garrett, A., Kelly, R., Gomez, R., Keller, J., Schatzberg, A. F., & Reiss, A. L. (2011). Aberrant Brain Activation During a Working Memory Task in Psychotic Major Depression. *American Journal of Psychiatry*, 168(2), 173–182.
<https://doi.org/10.1176/appi.ajp.2010.09121718>
- Gillis, M. M., Garcia, S., & Hampstead, B. M. (2016). Working memory contributes to the encoding of object location associations: Support for a 3-part model of object location memory. *Behavioural Brain Research*, 311, 192–200.
<https://doi.org/10.1016/j.bbr.2016.05.037>
- Goikolea, J. M., Dima, D., Landín-Romero, R., Torres, I., DelVecchio, G., Valentí, M., Amann, B. L., Bonnín, C. M., McKenna, P. J., Pomarol-Clotet, E., Frangou, S., & Vieta, E. (2019). Multimodal Brain Changes in First-Episode Mania: A Voxel-Based Morphometry, Functional Magnetic Resonance Imaging, and Connectivity Study. *Schizophrenia Bulletin*, 45(2), 464–473. <https://doi.org/10.1093/schbul/sby047>
- Gropman, A. L., Shattuck, K., Prust, M. J., Seltzer, R. R., Breeden, A. L., Hailu, A., Rigas, A., Hussain, R., & VanMeter, J. (2013). Altered neural activation in ornithine transcarbamylase deficiency during executive cognition: An fMRI study. *Human Brain Mapping*, 34(4), 753–761. <https://doi.org/10.1002/hbm.21470>
- Habel, U., Koch, K., Pauly, K., Kellermann, T., Reske, M., Backes, V., Seiferth, N. Y., Stöcker, T., Kircher, T., Amunts, K., Jon Shah, N., & Schneider, F. (2007). The influence of olfactory-induced negative emotion on verbal working memory: Individual

- differences in neurobehavioral findings. *Brain Research*, 1152, 158–170.
<https://doi.org/10.1016/j.brainres.2007.03.048>
- Harding, I. H., Corben, L. A., Storey, E., Egan, G. F., Stagnitti, M. R., Poudel, G. R., Delatycki, M. B., & Georgiou-Karistianis, N. (2016). Fronto-cerebellar dysfunction and dysconnectivity underlying cognition in friedreich ataxia: The IMAGE-FRDA study: Cognitive Networks in Friedreich Ataxia. *Human Brain Mapping*, 37(1), 338–350. <https://doi.org/10.1002/hbm.23034>
- Heinzel, S., Lorenz, R. C., Pelz, P., Heinz, A., Walter, H., Kathmann, N., Rapp, M. A., & Stelzel, C. (2016). Neural correlates of training and transfer effects in working memory in older adults. *NeuroImage*, 134, 236–249.
<https://doi.org/10.1016/j.neuroimage.2016.03.068>
- Honey, G. D., Sharma, T., Suckling, J., Giampietro, V., Soni, W., Williams, S. C. R., & Bullmore, E. T. (2003). The functional neuroanatomy of schizophrenic subsyndromes. *Psychological Medicine*, 33(6), 1007–1018.
<https://doi.org/10.1017/S0033291703007864>
- Honey, Garry D., Bullmore, E. T., & Sharma, T. (2000). Prolonged Reaction Time to a Verbal Working Memory Task Predicts Increased Power of Posterior Parietal Cortical Activation. *NeuroImage*, 12(5), 495–503.
<https://doi.org/10.1006/nimg.2000.0624>
- Huang, R.-R., Jia, B.-H., Xie, L., Ma, S.-H., Yin, J.-J., Sun, Z.-B., Le, H.-B., Xu, W.-C., Huang, J.-Z., & Luo, D.-X. (2016). Spatial working memory impairment in primary onset middle-age type 2 diabetes mellitus: An ethology and BOLD-fMRI study: Memory Impairment With Middle-Age Onset T2DM. *Journal of Magnetic Resonance Imaging*, 43(1), 75–87. <https://doi.org/10.1002/jmri.24967>
- Jiang, S., Yan, H., Chen, Q., Tian, L., Lu, T., Tan, H.-Y., Yan, J., & Zhang, D. (2015). Cerebral Inefficient Activation in Schizophrenia Patients and Their Unaffected Parents during the N-Back Working Memory Task: A Family fMRI Study. *PLOS ONE*, 10(8), e0135468. <https://doi.org/10.1371/journal.pone.0135468>
- Johannsen, L., Li, K. Z. H., Chechacz, M., Bibi, A., Kourtzi, Z., & Wing, A. M. (2013). Functional neuroimaging of the interference between working memory and the control of periodic ankle movement timing. *Neuropsychologia*, 51(11), 2142–2153. <https://doi.org/10.1016/j.neuropsychologia.2013.07.009>
- Jung, K., Friston, K. J., Pae, C., Choi, H. H., Tak, S., Choi, Y. K., Park, B., Park, C.-A., Cheong, C., & Park, H.-J. (2018). Effective connectivity during working memory and resting states: A DCM study. *NeuroImage*, 169, 485–495.
<https://doi.org/10.1016/j.neuroimage.2017.12.067>
- Kaminski, J., Gleich, T., Fukuda, Y., Katthagen, T., Gallinat, J., Heinz, A., & Schlagenhauf, F. (2020). Association of Cortical Glutamate and Working Memory Activation in Patients With Schizophrenia: A Multimodal Proton Magnetic Resonance Spectroscopy and Functional Magnetic Resonance Imaging Study. *Biological Psychiatry*, 87(3), 225–233. <https://doi.org/10.1016/j.biopsych.2019.07.011>
- Kim, J., Whyte, J., Wang, J., Rao, H., Tang, K. Z., & Detre, J. A. (2006). Continuous ASL perfusion fMRI investigation of higher cognition: Quantification of tonic CBF

- changes during sustained attention and working memory tasks. *NeuroImage*, 31(1), 376–385. <https://doi.org/10.1016/j.neuroimage.2005.11.035>
- King, T. Z., Na, S., & Mao, H. (2015). Neural Underpinnings of Working Memory in Adult Survivors of Childhood Brain Tumors. *Journal of the International Neuropsychological Society*, 21(7), 494–505. <https://doi.org/10.1017/S135561771500051X>
- Knops, A., Nuerk, H.-C., Fimm, B., Vohn, R., & Willmes, K. (2006). A special role for numbers in working memory? An fMRI study. *NeuroImage*, 29(1), 1–14. <https://doi.org/10.1016/j.neuroimage.2005.07.009>
- Koch, K., Pauly, K., Kellermann, T., Seiferth, N. Y., Reske, M., Backes, V., Stöcker, T., Shah, N. J., Amunts, K., Kircher, T., Schneider, F., & Habel, U. (2007). Gender differences in the cognitive control of emotion: An fMRI study. *Neuropsychologia*, 45(12), 2744–2754. <https://doi.org/10.1016/j.neuropsychologia.2007.04.012>
- Koppelstaetter, F., Poeppel, T. D., Siedentopf, C. M., Ischebeck, A., Verius, M., Haala, I., Mottaghy, F. M., Rhomberg, P., Golaszewski, S., Gotwald, T., Lorenz, I. H., Kolbitsch, C., Felber, S., & Krause, B. J. (2008). Does caffeine modulate verbal working memory processes? An fMRI study. *NeuroImage*, 39(1), 492–499. <https://doi.org/10.1016/j.neuroimage.2007.08.037>
- Korsnes, M. S., Lövdahl, H., Andersson, S., Björnerud, A., Due-Tønnesen, P., Endestad, T., & Malt, U. F. (2013). Working memory in recurrent brief depression: An fMRI pilot study. *Journal of Affective Disorders*, 149(1–3), 383–392. <https://doi.org/10.1016/j.jad.2013.02.017>
- Krug, A., Markov, V., Eggermann, T., Krach, S., Zerres, K., Stöcker, T., Shah, N. J., Schneider, F., Nöthen, M. M., Treutlein, J., Rietschel, M., & Kircher, T. (2008). Genetic variation in the schizophrenia-risk gene neuregulin1 correlates with differences in frontal brain activation in a working memory task in healthy individuals. *NeuroImage*, 42(4), 1569–1576. <https://doi.org/10.1016/j.neuroimage.2008.05.058>
- Kumari, V., Aasen, I., Taylor, P., ffytche, D. H., Das, M., Barkataki, I., Goswami, S., O’Connell, P., Howlett, M., Williams, S. C. R., & Sharma, T. (2006). Neural dysfunction and violence in schizophrenia: An fMRI investigation. *Schizophrenia Research*, 84(1), 144–164. <https://doi.org/10.1016/j.schres.2006.02.017>
- Lahr, J., Minkova, L., Tabrizi, S. J., Stout, J. C., Klöppel, S., Scheller, E., & the TrackOn-HD Investigators. (2018). Working Memory-Related Effective Connectivity in Huntington’s Disease Patients. *Frontiers in Neurology*, 9, 370. <https://doi.org/10.3389/fneur.2018.00370>
- Lamp, G., Alexander, B., Laycock, R., Crewther, D. P., & Crewther, S. G. (2016). Mapping of the Underlying Neural Mechanisms of Maintenance and Manipulation in Visuo-Spatial Working Memory Using An n-back Mental Rotation Task: A Functional Magnetic Resonance Imaging Study. *Frontiers in Behavioral Neuroscience*, 10. <https://doi.org/10.3389/fnbeh.2016.00087>
- Lee, T.-W., Liu, H.-L., Wai, Y.-Y., Ko, H.-J., & Lee, S.-H. (2013). Abnormal neural activity in partially remitted late-onset depression: An fMRI study of one-back working

- memory task. *Psychiatry Research: Neuroimaging*, 213(2), 133–141.
<https://doi.org/10.1016/j.psychresns.2012.04.010>
- Leung, A. W. S., & Alain, C. (2011). Working memory load modulates the auditory “What” and “Where” neural networks. *NeuroImage*, 55(3), 1260–1269.
<https://doi.org/10.1016/j.neuroimage.2010.12.055>
- Li, L., Men, W.-W., Chang, Y.-K., Fan, M.-X., Ji, L., & Wei, G.-X. (2014). Acute Aerobic Exercise Increases Cortical Activity during Working Memory: A Functional MRI Study in Female College Students. *PLoS ONE*, 9(6), e99222.
<https://doi.org/10.1371/journal.pone.0099222>
- Li, L., Zhang, S., Cui, J., Chen, L.-Z., Wang, X., Fan, M., & Wei, G.-X. (2019). Fitness-Dependent Effect of Acute Aerobic Exercise on Executive Function. *Frontiers in Physiology*, 10, 902. <https://doi.org/10.3389/fphys.2019.00902>
- Li, X., Yi, Z., Lv, Q., Chu, M., Hu, H., Wang, J., Zhang, J., Cheung, E. E. F., & Chan, R. C. K. (2019). Clinical utility of the dual n-back task in schizophrenia: A functional imaging approach. *Psychiatry Research: Neuroimaging*, 284, 37–44.
<https://doi.org/10.1016/j.psychresns.2019.01.002>
- Lim, H.-K., Juh, R., Pae, C.-U., Lee, B.-T., Yoo, S.-S., Ryu, S.-H., Kwak, K.-R., Lee, C., & Lee, C.-U. (2008). Altered Verbal Working Memory Process in Patients with Alzheimer’s Disease. *Neuropsychobiology*, 57(4), 181–187.
<https://doi.org/10.1159/000147471>
- Luo, Y., Qin, S., Fernández, G., Zhang, Y., Klumpers, F., & Li, H. (2014). Emotion perception and executive control interact in the salience network during emotionally charged working memory processing: Examination of Neural Mechanisms on Processing of Emotional WM. *Human Brain Mapping*, 35(11), 5606–5616.
<https://doi.org/10.1002/hbm.22573>
- Malisza, K. L., Allman, A.-A., Shiloff, D., Jakobson, L., Longstaffe, S., & Chudley, A. E. (2005). Evaluation of Spatial Working Memory Function in Children and Adults with Fetal Alcohol Spectrum Disorders: A Functional Magnetic Resonance Imaging Study. *Pediatric Research*, 58(6), 1150–1157.
<https://doi.org/10.1203/01.pdr.0000185479.92484.a1>
- Marquand, A. F., Mourão-Miranda, J., Brammer, M. J., Cleare, A. J., & Fu, C. H. Y. (2008). Neuroanatomy of verbal working memory as a diagnostic biomarker for depression: *NeuroReport*, 19(15), 1507–1511.
<https://doi.org/10.1097/WNR.0b013e328310425e>
- Matsuo, K., Glahn, D. C., Peluso, M. A. M., Hatch, J. P., Monkul, E. S., Najt, P., Sanches, M., Zamarripa, F., Li, J., Lancaster, J. L., Fox, P. T., Gao, J.-H., & Soares, J. C. (2007). Prefrontal hyperactivation during working memory task in untreated individuals with major depressive disorder. *Molecular Psychiatry*, 12(2), 158–166.
<https://doi.org/10.1038/sj.mp.4001894>
- McAllister, T. W., Saykin, A. J., Flashman, L. A., Sparling, M. B., Johnson, S. C., Guerin, S. J., Mamourian, A. C., Weaver, J. B., & Yanofsky, N. (1999). Brain activation during working memory 1 month after mild traumatic brain injury: A functional MRI study. *Neurology*, 53(6), 1300–1300. <https://doi.org/10.1212/WNL.53.6.1300>

- McGeown, W. J., Shanks, M. F., & Venneri, A. (2008). Prolonged cholinergic enrichment influences regional cortical activation in early Alzheimer's disease. *Neuropsychiatric Disease and Treatment*, 465. <https://doi.org/10.2147/NDT.S2461>
- Meisenzahl, E. M., Scheuerecker, J., Zipse, M., Ufer, S., Wiesmann, M., Frodl, T., Koutsouleris, N., Zetzsche, T., Schmitt, G., Riedel, M., Spellmann, I., Dehning, S., Linn, J., Brückmann, H., & Möller, H. J. (2006). Effects of treatment with the atypical neuroleptic quetiapine on working memory function: A functional MRI follow-up investigation. *European Archives of Psychiatry and Clinical Neuroscience*, 256(8), 522–531. <https://doi.org/10.1007/s00406-006-0687-x>
- Migo, E. M., Mitterschiffthaler, M., O'Daly, O., Dawson, G. R., Dourish, C. T., Craig, K. J., Simmons, A., Wilcock, G. K., McCulloch, E., Jackson, S. H. D., Kopelman, M. D., Williams, S. C. R., & Morris, R. G. (2015). Alterations in working memory networks in amnesic mild cognitive impairment. *Aging, Neuropsychology, and Cognition*, 22(1), 106–127. <https://doi.org/10.1080/13825585.2014.894958>
- Miró-Padilla, A., Bueichekú, E., Ventura-Campos, N., Flores-Compañ, M.-J., Parcet, M. A., & Ávila, C. (2019). Long-term brain effects of N-back training: An fMRI study. *Brain Imaging and Behavior*, 13(4), 1115–1127. <https://doi.org/10.1007/s11682-018-9925-x>
- Monks, P. J., Thompson, J. M., Bullmore, E. T., Suckling, J., Brammer, M. J., Williams, S. C., Simmons, A., Giles, N., Lloyd, A. J., Louise Harrison, C., Seal, M., Murray, R. M., Nicol Ferrier, I., Young, A. H., & Curtis, V. A. (2004). A functional MRI study of working memory task in euthymic bipolar disorder: Evidence for task-specific dysfunction. *Bipolar Disorders*, 6(6), 550–564. <https://doi.org/10.1111/j.1399-5618.2004.00147.x>
- Nebel, K., Wiese, H., Stude, P., de Greiff, A., Diener, H.-C., & Keidel, M. (2005). On the neural basis of focused and divided attention. *Cognitive Brain Research*, 25(3), 760–776. <https://doi.org/10.1016/j.cogbrainres.2005.09.011>
- Ogg, R. J., Zou, P., Allen, D. N., Hutchins, S. B., Dutkiewicz, R. M., & Mulhern, R. K. (2008). Neural correlates of a clinical continuous performance test. *Magnetic Resonance Imaging*, 26(4), 504–512. <https://doi.org/10.1016/j.mri.2007.09.004>
- Park, J.-W., Kim, Y.-T., Yun, B.-J., Jin, S.-U., Lee, S.-H., Ahn, S.-H., Min, Y., Jung, T.-D., Lee, H. J., & Chang, Y. (2016). Stereoscopic 3D objects evoke stronger saliency for nonverbal working memory: An fMRI study. *International Journal of Imaging Systems and Technology*, 26(1), 76–84. <https://doi.org/10.1002/ima.22159>
- Pfefferbaum, A., Desmond, J. E., Galloway, C., Menon, V., Glover, G. H., & Sullivan, E. V. (2001). Reorganization of Frontal Systems Used by Alcoholics for Spatial Working Memory: An fMRI Study. *NeuroImage*, 14(1), 7–20. <https://doi.org/10.1006/nimg.2001.0785>
- Philip, N. S., Sweet, L. H., Tyrka, A. R., Carpenter, S. L., Albright, S. E., Price, L. H., & Carpenter, L. L. (2016). Exposure to childhood trauma is associated with altered n-back activation and performance in healthy adults: Implications for a commonly used working memory task. *Brain Imaging and Behavior*, 10(1), 124–135. <https://doi.org/10.1007/s11682-015-9373-9>

- Pomarol-Clotet, E., Salvador, R., Sarró, S., Gomar, J., Vila, F., Martínez, Á., Guerrero, A., Ortiz-Gil, J., Sans-Sansa, B., Capdevila, A., Cebamanos, J. M., & McKenna, P. J. (2008). Failure to deactivate in the prefrontal cortex in schizophrenia: Dysfunction of the default mode network? *Psychological Medicine*, 38(8), 1185–1193. <https://doi.org/10.1017/S0033291708003565>
- Qin, S., Hermans, E. J., van Marle, H. J. F., Luo, J., & Fernández, G. (2009). Acute Psychological Stress Reduces Working Memory-Related Activity in the Dorsolateral Prefrontal Cortex. *Biological Psychiatry*, 66(1), 25–32. <https://doi.org/10.1016/j.biopsych.2009.03.006>
- Ragland, J. D., Turetsky, B. I., Gur, R. C., Gunning-Dixon, F., Turner, T., Schroeder, L., Chan, R., & Gur, R. E. (2002). Working memory for complex figures: An fMRI comparison of letter and fractal n-back tasks. *Neuropsychology*, 16(3), 370–379.
- Ricciardi, E., Bonino, D., Gentili, C., Sani, L., Pietrini, P., & Vecchi, T. (2006). Neural correlates of spatial working memory in humans: A functional magnetic resonance imaging study comparing visual and tactile processes. *Neuroscience*, 139(1), 339–349. <https://doi.org/10.1016/j.neuroscience.2005.08.045>
- Richter, S., Gorny, X., Machts, J., Behnisch, G., Wüstenberg, T., Herbort, M. C., Münte, T. F., Seidenbecher, C. I., & Schott, B. H. (2013). Effects of AKAP5 Pro100Leu Genotype on Working Memory for Emotional Stimuli. *PLoS ONE*, 8(1), e55613. <https://doi.org/10.1371/journal.pone.0055613>
- Rodríguez-Cano, E., Sarró, S., Monté, G. C., Maristany, T., Salvador, R., McKenna, P. J., & Pomarol-Clotet, E. (2014). Evidence for structural and functional abnormality in the subgenual anterior cingulate cortex in major depressive disorder. *Psychological Medicine*, 44(15), 3263–3273. <https://doi.org/10.1017/S0033291714000841>
- Rodríguez-Cano, Elena, Alonso-Lana, S., Sarró, S., Fernández-Corcuera, P., Goikolea, J. M., Vieta, E., Maristany, T., Salvador, R., McKenna, P. J., & Pomarol-Clotet, E. (2017). Differential failure to deactivate the default mode network in unipolar and bipolar depression. *Bipolar Disorders*, 19(5), 386–395. <https://doi.org/10.1111/bdi.12517>
- Salavert, J., Ramos-Quiroga, J. A., Moreno-Alcázar, A., Caseras, X., Palomar, G., Radua, J., Bosch, R., Salvador, R., McKenna, P. J., Casas, M., & Pomarol-Clotet, E. (2018). Functional Imaging Changes in the Medial Prefrontal Cortex in Adult ADHD. *Journal of Attention Disorders*, 22(7), 679–693. <https://doi.org/10.1177/1087054715611492>
- Sánchez-Carrión, R., Gómez, P. V., Junqué, C., Fernández-Espejo, D., Falcon, C., Bargalló, N., Roig-Rovira, T., Enseñat-Cantalops, A., & Bernabeu, M. (2008). Frontal Hypoactivation on Functional Magnetic Resonance Imaging in Working Memory after Severe Diffuse Traumatic Brain Injury. *Journal of Neurotrauma*, 25(5), 479–494. <https://doi.org/10.1089/neu.2007.0417>
- Sapara, A., ffytche, D. H., Birchwood, M., Cooke, M. A., Fannon, D., Williams, S. C. R., Kuipers, E., & Kumari, V. (2014). Preservation and compensation: The functional neuroanatomy of insight and working memory in schizophrenia. *Schizophrenia Research*, 152(1), 201–209. <https://doi.org/10.1016/j.schres.2013.11.026>

- Scheller, E., Peter, J., Schumacher, L. V., Lahr, J., Mader, I., Kaller, C. P., & Klöppel, S. (2017). APOE moderates compensatory recruitment of neuronal resources during working memory processing in healthy older adults. *Neurobiology of Aging*, 56, 127–137. <https://doi.org/10.1016/j.neurobiolaging.2017.04.015>
- Scheuerecker, J., Ufer, S., Zipse, M., Frodl, T., Koutsouleris, N., Zetzsche, T., Wiesmann, M., Albrecht, J., Brückmann, H., Schmitt, G., Möller, H.-J., & Meisenzahl, E. M. (2008). Cerebral changes and cognitive dysfunctions in medication-free schizophrenia – An fMRI study. *Journal of Psychiatric Research*, 42(6), 469–476. <https://doi.org/10.1016/j.jpsychires.2007.04.001>
- Schlagenhauf, F., Wüstenberg, T., Schmack, K., Dinges, M., Wrase, J., Koslowski, M., Kienast, T., Bauer, M., Gallinat, J., Juckel, G., & Heinz, A. (2008). Switching schizophrenia patients from typical neuroleptics to olanzapine: Effects on BOLD response during attention and working memory. *European Neuropsychopharmacology*, 18(8), 589–599. <https://doi.org/10.1016/j.euroneuro.2008.04.013>
- Schmidt, C., Collette, F., Reichert, C. F., Maire, M., Vandewalle, G., Peigneux, P., & Cajochen, C. (2015). Pushing the Limits: Chronotype and Time of Day Modulate Working Memory-Dependent Cerebral Activity. *Frontiers in Neurology*, 6. <https://doi.org/10.3389/fneur.2015.00199>
- Schneider, F., Habel, U., Reske, M., Kellermann, T., Stöcker, T., Shah, N. J., Zilles, K., Braus, D. F., Schmitt, A., Schlösser, R., Wagner, M., Frommann, I., Kircher, T., Rapp, A., Meisenzahl, E., Ufer, S., Ruhrmann, S., Thienel, R., Sauer, H., ... Gaebel, W. (2007). Neural correlates of working memory dysfunction in first-episode schizophrenia patients: An fMRI multi-center study. *Schizophrenia Research*, 89(1–3), 198–210. <https://doi.org/10.1016/j.schres.2006.07.021>
- Schneiders, J. A., Opitz, B., Krick, C. M., & Mecklinger, A. (2011). Separating Intra-Modal and Across-Modal Training Effects in Visual Working Memory: An fMRI Investigation. *Cerebral Cortex*, 21(11), 2555–2564. <https://doi.org/10.1093/cercor/bhr037>
- Seo, J., Kim, S.-H., Kim, Y.-T., Song, H., Lee, J., Kim, S.-H., Han, S. W., Nam, E. J., Kim, S.-K., Lee, H. J., Lee, S.-J., & Chang, Y. (2012). Working Memory Impairment in Fibromyalgia Patients Associated with Altered Frontoparietal Memory Network. *PLoS ONE*, 7(6), e37808. <https://doi.org/10.1371/journal.pone.0037808>
- Seo, J., Lee, B.-K., Jin, S.-U., Park, J. W., Kim, Y.-T., Ryeom, H.-K., Lee, J., Suh, K. J., Kim, S. H., Park, S.-J., Jeong, K. S., Ham, J.-O., Kim, Y., & Chang, Y. (2014). Lead-Induced Impairments in the Neural Processes Related to Working Memory Function. *PLoS ONE*, 9(8), e105308. <https://doi.org/10.1371/journal.pone.0105308>
- Shen, L., Hu, X., Yacoub, E., & Ugurbil, K. (1999). Neural correlates of visual form and visual spatial processing. *Human Brain Mapping*, 8(1), 60–71.
- Smits, M., Dippel, D. W. J., Houston, G. C., Wielopolski, P. A., Koudstaal, P. J., Hunink, M. G. M., & van der Lugt, A. (2009). Postconcussion syndrome after minor head injury: Brain activation of working memory and attention. *Human Brain Mapping*, 30(9), 2789–2803. <https://doi.org/10.1002/hbm.20709>

- Spreng, R. N., DuPre, E., Selarka, D., Garcia, J., Gojkovic, S., Mildner, J., Luh, W.-M., & Turner, G. R. (2014). Goal-Congruent Default Network Activity Facilitates Cognitive Control. *Journal of Neuroscience*, *34*(42), 14108–14114. <https://doi.org/10.1523/JNEUROSCI.2815-14.2014>
- Thomas, R. J., Rosen, B. R., Stern, C. E., Weiss, J. W., & Kwong, K. K. (2005). Functional imaging of working memory in obstructive sleep-disordered breathing. *Journal of Applied Physiology*, *98*(6), 2226–2234. <https://doi.org/10.1152/japplphysiol.01225.2004>
- Thornton, M. A., & Conway, A. R. A. (2013). Working memory for social information: Chunking or domain-specific buffer? *NeuroImage*, *70*, 233–239. <https://doi.org/10.1016/j.neuroimage.2012.12.063>
- Vacchi, L., Rocca, M. A., Meani, A., Rodegher, M., Martinelli, V., Comi, G., Falini, A., & Filippi, M. (2017). Working memory network dysfunction in relapse-onset multiple sclerosis phenotypes: A clinical-imaging evaluation. *Multiple Sclerosis Journal*, *23*(4), 577–587. <https://doi.org/10.1177/1352458516656809>
- van der Horn, H. J., Liemburg, E. J., Scheenen, M. E., de Koning, M. E., Spikman, J. M., & van der Naalt, J. (2016). Post-concussive complaints after mild traumatic brain injury associated with altered brain networks during working memory performance. *Brain Imaging and Behavior*, *10*(4), 1243–1253. <https://doi.org/10.1007/s11682-015-9489-y>
- Waiter, G. D., Deary, I. J., Staff, R. T., Murray, A. D., Fox, H. C., Starr, J. M., & Whalley, L. J. (2009). Exploring possible neural mechanisms of intelligence differences using processing speed and working memory tasks: An fMRI study. *Intelligence*, *37*(2), 199–206. <https://doi.org/10.1016/j.intell.2008.09.008>
- Walitt, B., Čeko, M., Khatiwada, M., Gracely, J. L., Rayhan, R., VanMeter, J. W., & Gracely, R. H. (2016). Characterizing “fibrofog”: Subjective appraisal, objective performance, and task-related brain activity during a working memory task. *NeuroImage: Clinical*, *11*, 173–180. <https://doi.org/10.1016/j.nicl.2016.01.021>
- Wesley, M. J., Lile, J. A., Fillmore, M. T., & Porrino, L. J. (2017). Neurophysiological capacity in a working memory task differentiates dependent from nondependent heavy drinkers and controls. *Drug and Alcohol Dependence*, *175*, 24–35. <https://doi.org/10.1016/j.drugalcdep.2017.01.029>
- Wishart, H. A., Saykin, A. J., Rabin, L. A., Santulli, R. B., Flashman, L. A., Guerin, S. J., Mamourian, A. C., Belloni, D. R., Rhodes, C. H., & McAllister, T. W. (2006). Increased Brain Activation During Working Memory in Cognitively Intact Adults With the APOE ε4 Allele. *American Journal of Psychiatry*, *163*(9), 1603–1610. <https://doi.org/10.1176/ajp.2006.163.9.1603>
- Wu, S., Wang, H., Chen, C., Zou, J., Huang, H., Li, P., Zhao, Y., Xu, Q., Zhang, L., Wang, H., Pandit, S., Dahal, S., Chen, J., Zhou, Y., Jiang, T., & Wang, G. (2017). Task Performance Modulates Functional Connectivity Involving the Dorsolateral Prefrontal Cortex in Patients with Schizophrenia. *Frontiers in Psychology*, *8*. <https://doi.org/10.3389/fpsyg.2017.00056>

- Yan, X., Zhang, J., Gong, Q., & Weng, X. (2011). Adaptive influence of long term high altitude residence on spatial working memory: An fMRI study. *Brain and Cognition*, 77(1), 53–59. <https://doi.org/10.1016/j.bandc.2011.06.002>
- Yang, X., Zhang, X., Yang, Y., & Lin, N. (2018). How context features modulate the involvement of the working memory system during discourse comprehension. *Neuropsychologia*, 111, 36–44. <https://doi.org/10.1016/j.neuropsychologia.2018.01.010>
- Yoo, S.-S., Choi, B.-G., Juh, R.-H., Park, J.-M., Pae, C.-U., Kim, J.-J., Lee, S.-J., Lee, C., Paik, I.-H., Lee, C.-U., & Adkinson, N. F. (2005). WORKING MEMORY PROCESSING OF FACIAL IMAGES IN SCHIZOPHRENIA: FMRI INVESTIGATION. *International Journal of Neuroscience*, 115(3), 351–366. <https://doi.org/10.1080/00207450590520957>
- Yoo, S.-S., Paralkar, G., & Panych, L. P. (2004). NEURAL SUBSTRATES ASSOCIATED WITH THE CONCURRENT PERFORMANCE OF DUAL WORKING MEMORY TASKS. *International Journal of Neuroscience*, 114(6), 613–631. <https://doi.org/10.1080/00207450490430561>
- Ziemus, B., Baumann, O., Luerding, R., Schlosser, R., Schuierer, G., Bogdahn, U., & Greenlee, M. (2007). Impaired working-memory after cerebellar infarcts paralleled by changes in BOLD signal of a cortico-cerebellar circuit. *Neuropsychologia*, 45(9), 2016–2024. <https://doi.org/10.1016/j.neuropsychologia.2007.02.012>