Working Memory Updating Training Promotes Plasticity & Behavioural Gains:

A Systematic Review & Meta-Analysis

Katerina Pappa[*1], Viveka Biswas[2], Kristin Flegal[2], Jonathan Evans[1], Satu Baylan[1]*

[1] Institute of Health and Wellbeing, University of Glasgow, [2] Institute of Neuroscience and Psychology, University of Glasgow

* **Corresponding author:** Imaging Centre of Excellence, Queen Elizabeth University Hospital, Glasgow, G51 4TF, **Tel:** 0141 451 6863, **Email:** a.pappa.1@research.gla.ac.uk.

Abstract

Aims Recent reviews yield contradictory findings regarding the efficacy of working memory training and transfer to untrained tasks. We reviewed working memory updating (WMU) training studies and examined cognitive and neural outcomes on training and transfer tasks. **Methods** Database searches for adult brain imaging studies of WMU training were conducted. Training-induced neural changes were assessed qualitatively, and meta-analyses were performed on behavioural training and transfer effects. **Results** A large behavioural training effect was found for WMU training groups compared to control groups. There was a moderate near transfer effect on tasks in the same cognitive domain, and a non-significant effect for far transfer to other cognitive domains. Functional neuroimaging changes for WMU training tasks revealed consistent frontoparietal activity decreases while both decreases and increases were

found for subcortical regions. **Conclusions** WMU training promotes plasticity and has potential applications in optimizing interventions for neurological populations. Future research should focus on the mechanisms and factors underlying plasticity and generalisation of training gains.

Keywords: plasticity, learning, working memory updating, cognitive training, transfer, neuroimaging.

1. Introduction

In cognitive neuroscience, an emerging research area concerns experience-induced changes in brain structure and function, referred to as *plasticity*. Plasticity has traditionally been defined as the capacity of the brain to adjust in response to environmental changes and it is considered to mediate acquisition of knowledge, skill, and repair after injury (1). For example, plasticity is seen as the restoration and compensation of the neural system following a brain injury. Similarly, following training on a cognitive task, the neural system's response to the training – i.e. the improved cognitive performance and the structural changes in the brain's system– are also considered indications of plasticity (2).

Structural changes can be direct, including neurogenesis (formation of new neurons), gliogenesis (formation of new glial cells), dendritic or axonal growth, as well as indirect changes to the system's function, such as angiogenesis (formation of new blood vessels). Both direct and indirect changes are considered structural changes in the overall neural system. Within Lövdén and colleagues' theoretical framework for plasticity (2), these structural changes in the system can be measured as changes in: 1. the structure of the brain, e.g. changes in gray matter volume and white matter microstructure, 2. the molecular scale, e.g. changes in receptor density and 3. the function of the brain, e.g. changes in activation patterns. Therefore, signs of plasticity are measurable with neuroimaging methodologies such as structural and functional Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), Arterial Spin Labelling (ASL), and Diffusion Tensor Imaging (DTI).

Lövdén et al. (2) further defined the term *flexibility* as the neural system's existing ability to adapt effectively to environmental demands and utilise the neural processes necessary for performing a given task. This is in contrast to the concept of plasticity, defined as the system's response to meeting prolonged changes in environmental demands through learning, and structural alterations, which subsequently produces a change in the pre-existing adaptive ability. Lövdén et al. (2) explained this by theorising that a mismatch between functional

″supply″ (i.e., neural resources) and environmental ″demands″ (e.g., a continuously challenging cognitive task) is a necessary condition for plasticity to occur.

Working memory (WM) refers to a system that is essential for the maintenance and manipulation of information in order to successfully perform complex cognitive tasks such as learning and language comprehension (3). The classic WM model consists of three components: two slave systems (i.e., the phonological loop handling speech-based information and the visuospatial sketchpad manipulating visual images) and the central executive, an attentional control system responsible for the regulation of cognitive processes, i.e., executive functions (3, 4). It has been argued that executive functioning depends upon three processes: 1. shifting attention between tasks and active representations, 2. inhibition of automatic responses and irrelevant information; 3. working memory updating (WMU), i.e., modifying the content of WM according to incoming information (5). Miyake et al. (4) proposed that these executive functions are correlated with each other but are also distinct from one another.

Neuropsychological and neuroimaging studies have established the reliance of these executive functions upon the prefrontal cortex (PFC) and parietal regions, in addition to PFC interactions with subcortical structures such as basal ganglia and thalamus (5-12). Key regions forming the neural basis of WM comprise the mid-ventrolateral PFC (VLPFC) including the inferior frontal gyrus (IFG) pars triangularis, and IFG pars opercularis; dorsolateral PFC (DLPFC); precentral gyrus (preCG); posterior parietal cortex (PPC) including the superior parietal lobule (SPL) and inferior parietal lobule (IPL); temporo-parietal junction (TPJ) (5, 12); and subcortical regions such as the basal ganglia involving the striatum (caudate nucleus and putamen) (7, 11, 12).

There is evidence to suggest that training cognitive processes, including WM executive functions, produces plastic changes (13, 14) demonstrated by improved cognitive performance and neural changes. Cognitive training research, however, frequently faces criticisms that the cognitive improvement is limited to the task being trained, i.e., criterion task,

and does not generalise (or transfer) to other untrained tasks (15-18). Similarly, in cognitive training studies including neuroimaging outcome measures, there is no consensus regarding the pattern of training-induced functional and structural changes (19, 20). There have been a number of meta-analyses and systematic reviews of cognitive outcomes (16-18, 21), including some examining both cognitive and neural outcomes, following WM training (13, 14, 19, 20). Despite the increasing interest in WM training, different studies have presented contradictory findings concerning key issues (18).

1.1 Cognitive Performance Changes following WM training

Previous meta-analyses evaluating the efficacy of WM training have concentrated on: 1) transfer of training gains to untrained tasks, and degree of similarity to the trained criterion for untrained tasks in which this is observed (i.e., near or far transfer), 2) features of the training intervention, with the type of control group, age of the participants, training dose and specific training task most examined.

Different meta-analytic reviews have arrived at conflicting conclusions, with some authors (21) finding evidence for far transfer (to more general cognitive domains) after WM n-back training and others concluding there are data to support near transfer effects (within the same cognitive domain) but very small or no evidence of far transfer (16-18). Inconsistencies regarding the employment of an active or passive control group have also been reported, with some authors determining the type of control group does not affect the size of the transfer effect (18, 21, 22), and others concluding there is no evidence of far transfer when comparing training groups against active control groups. The latter finding suggests that the transfer effect is overestimated when employing passive control groups (15, 17). In a theoretical review, Von Bastian and Oberauer (23) state that more training sessions lead to a larger training effect while no consensus is reached regarding the most optimal spacing and scheduling of training sessions. The review concludes that the effect of training declines with age and suggests a lack of consistency in the evidence favouring training protocols with adaptive task difficulty.

1.2 Neural Changes following WM training

Functional activation increases in practice-related neuroimaging studies are explained as added recruitment of brain regions or as response strengthening within a cortical region (24) and is usually seen after practice on motor or sensory tasks. Functional activation decreases, on the other hand, are explained as increased efficiency, indicating that fewer neurons needing to fire when responding to a stimulus (24). This is interpreted as a robust and efficient neural representation and is usually observed after training higher cognitive processes such as WM (24). Reorganisation of activation is commonly observed after practice and two types can be distinguished: 1) redistribution of functional activations and 2) functional reorganisation of activation (24).

Neural changes induced by WM training have been observed in healthy young and older adults in fronto-parietal cortical regions and subcortical regions, e.g., the striatal system involving caudate nucleus and putamen; however, the direction of these changes after training is inconsistent (20). A comprehensive fMRI meta-analysis by Salmi et al. (12) examined the neural changes following all types of WM training and provided valuable insight into key issues including: 1. features of the neural networks exhibiting training-related modulations; 2. dynamic changes of the functional activity patterns when comparing training paradigms of shorter and longer duration and 3. patterns of training-related neural modulation in transfer tasks.

The meta-analysis concluded that activity decreases after WM training were more often reported and more consistent in the DLPFC area, while increases were reported less frequently and related to areas involved in the salience network and dorsal attention network as well as striatum and thalamus. The same review suggested that training-related neural changes are manifested in existing core WM networks including the dorsal attention and salience networks, the DLPFC and striatum, rather than recruitment of new networks following training (i.e. redistribution of functional activations within the same network) (12). This observation proposes a direct relationship between a region's involvement in WM and training-

related modulation in that region. Another interesting finding is the consistency of fronto-parietal activations and modulations in studies of any training duration, while activity modulations in the DLPFC and striatum were only evident in longer training protocols (i.e. more than two weeks). Overall, training-related activity pattern changes in transfer tasks have not been examined as extensively as for the trained criterion task. However, a meta-analysis of the training-related neural modulation for untrained transfer tasks revealed increases in the striatum and IFG and decreases in the DLPFC suggesting the fronto-striatal system mediates transfer of WM training (12).

In contrast to functional activity outcome measurements, only a handful of studies to date have explored changes in functional connectivity after WM training, making it difficult to draw confident conclusions, though the studies report increases in fronto-parietal networks overall (25, 26).

Alterations in brain structure as a result of training may involve changes in grey matter volume or cortical thickness in task-relevant regions and changes in white matter volume and microstructure, predominantly measured as fractional anisotropy (FA) using DTI (27, 28). FA is thought to be modulated by myelination and is considered an indication of structural connection strength, axon diameter and density (28). Few studies to date have focused on structural changes after WM training. Nevertheless, one study reported reduced grey matter in frontal and parietal cortices (29) and another found both cortical thicknesses increases and decreases in frontal areas (17, 30). Structural connectivity increases in the fronto-parietal network have also been reported following WM training (31, 32).

1.3 The current review

The majority of published reviews to date are broad and include studies with a plethora of WM training tasks involving various processes and tapping into multiple executive functions such as shifting and inhibition as well as WMU (4). Consequently, this variability has made it difficult to draw consistent conclusions on the efficacy of WM training (18, 20). In our review we focus solely on the updating process of WM to achieve greater homogeneity of the process being

trained, regardless of modality and task parameters. For example, even though the recent fMRI meta-analysis by Salmi et al. (12) provides a comprehensive overview of the neural modulations following WM training, in addition to its basis on a large data sample, our review examines process-specific outcomes by focusing on the effects of WMU training exclusively.

In the cognitive training literature, the updating process of WM has been examined using different task paradigms such as memory updating and n-back. A working memory updating task paradigm requires participants to store and update incoming stimuli such as letters, digits or spatial locations, while performing a series of operations, e.g. spatial location changes, arithmetic operations (33). Another WMU paradigm involves the n-back task where participants are required to store and update the last n elements, e.g. numbers, letters, spatial locations; and then decide if the most recently presented item matches the one shown n steps back (33). The n-back task taxes various cognitive processes simultaneously, aside from updating, such as encoding, monitoring and maintenance (34). It is a very frequently used paradigm in WM training studies (35) due to its usefulness in experimental research (34).

A study by Schmiedek et al. (33) reported high latent correlations of n-back and memory updating tasks and further concluded that both paradigms provide good measurements of WM. Linares et al. (35) investigated the transfer effects following WM training comparing a memory updating training group and an n-back training group against an active control group. Both training groups improved their performance on their respective trained task, but none exhibited near or far transfer of learning. Furthermore, even though both paradigms involved the WMU process; performance gains on the memory updating task did not lead to gains in the n-back task and vice versa, suggesting the tasks vary in other cognitive processes. Even though the memory updating and n-back tasks are not alike in every way, and each involves additional distinct cognitive processes, nonetheless they both tap into the WMU process (35). Consequently, the current review includes studies using both training paradigms, especially since it is the first to focus on the process of updating exclusively.

The aim of this review is to examine the cognitive and neural outcomes of WMU training and transfer to untrained tasks. Meta-analyses on cognitive outcomes in the reviewed studies that assess task-based functional neuroimaging data is undertaken to further investigate the training-related effects in adults. The cognitive outcomes focus on the training and transfer effect sizes while the neural outcomes report on the changes following WMU training in terms of functional activation as well as functional connectivity and structural imaging measures for both training and transfer tasks.

2. Methods

This work was prepared in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (36) and was registered on PROSPERO, the international prospective register of systematic reviews, (ID number: CRD42019120234).

2.1 Database Search and Study selection

A comprehensive search was conducted to identify studies that investigated cognitive and neuroimaging outcomes following WMU training in adults. Before proceeding with the final database searches, we repeatedly tested the sensitivity of a combination of key words and Medical Subject Headings (MeSH) to make sure our searches would be comprehensive and rigorous. We used the fMRI study by Dahlin et al. (37) as an exemplar to inspect and confirm the search relevance in the different database searches. We further noted the keywords and MeSH terms listed for relevant studies in the different databases and tried to incorporate them in our search terms. When these didn't capture the exemplar study, the search terms were further refined. Once we were confident our search strategy was fitting, rigorous and that the exemplar study was identified in all databases, we then proceeded with the final search.

The studies were published up to and inclusive of 28th January 2019 in the first instance. An updated search was conducted for publications between January 2019 and 13th June 2020. The articles were sought from Ovid EMBASE, Ovid MEDLINE, PsycINFO, CINAHL, Scopus and Cochrane Library electronic databases consisting of the following MeSH and keyword

search terms: 1. working memory OR executive function OR (“Working memory” adj5 train*) OR (“Working memory” adj5 updat*) OR (n-back adj5 train*), 2. training OR intervention OR remediation), 3. functional magnetic resonance imaging/ OR (FMRI OR PET OR MRI OR “resting state”). These search terms were then combined using a boolean operator “AND”. Our search strategy was pre-registered on PROSPERO. Only peer-reviewed journals and articles written in English were included. Titles and abstracts were screened independently by two reviewers (KP & VB) while full-texts were screened against inclusion criteria and when discrepancies occurred, a third reviewer was consulted (SB).

2.2 Eligibility Criteria

This systematic review included studies on adults over the age of 18. We included healthy participants as well as adults with neurological conditions, while psychiatric samples were excluded. Any type of experimental research design, i.e., both non-randomised and randomised controlled trials, cross-over trials and single-case studies were included. The studies included any type of control group (CG), i.e., active CG, passive CG, and no CG. We included studies of any duration which trained the process specific to WMU regardless of training modality. Studies that used a WM training regime that was not specific to the WMU process were excluded, as well as other cognitive training unrelated to WM or multi-domain training. Our criteria in terms of the neuroimaging methodology were broad in that functional magnetic resonance imaging (fMRI), PET, ASL, structural imaging and functional connectivity studies were all of interest. We only included studies that conducted more than one neuroimaging session, i.e., before and after WMU training, regardless of the total number of imaging sessions that took place after WMU training had commenced.

2.3 Outcomes

Our primary outcomes included cognitive and neural changes as a result of WMU training. In both cases we concentrated on the trained task, i.e., criterion task, to examine the training effect. If studies assessed the transfer of training to untrained tasks, then the transfer effect (cognitive and/or neural outcomes) was explored as a secondary outcome. The transfer

effects were further subdivided into near transfer (within the same cognitive domain) and far transfer (to other more general cognitive domains).

2.4 Data Extraction & Synthesis

We created and piloted a list of data extraction items under three categories. The first included study characteristics, i.e., sample size and demographics, study design, number of scanning sessions, type of neuroimaging outcome, description of the tasks performed during brain imaging as well as independent to the scanning sessions. The second category listed information on the WMU training protocol followed by each study, i.e., training task, type and modality, training duration (total number of sessions and duration per session), total hours of training and information on the control group. The final category contained information on the cognitive and neural outcomes separated in terms of the specific neuroimaging methodology utilized. Data on the effect of training and/or transfer were extracted separately for tasks assessed inside or outside the scanner. For both cognitive and neural outcomes, data on the group by time interaction together with significance level and F values were extracted if an ANOVA test was performed. Means and standard deviations (SDs) for each group pre and post training were also noted. We tried to extract data on the same statistical test across all studies to keep our data synthesis as homogeneous and unbiased as possible.

2.5 Quality Assessment

Methodological quality of studies was assessed using the Physiotherapy Evidence Database Rating Scale (PEDro-P) scale (38). This tool was chosen as it is the primary scale used in the NeuroRehab Evidence Resource (NeuroBITE, previously PsychBITE) to evaluate methodological quality for trials of cognitive, behavioural and other treatments. NeuroBITE offers an online extensive training program and scoring guidelines on the PEDro-P scale[1]. The PEDro-P scale contains eleven items relating to the external and internal validity of the study. The first item is related to external validity and is not included in the overall score, the maximum quality assessment score on the scale is 10. A rating of 1 is awarded for each item

[1] http://www.neurorehab-evidence.com/web/cms/content/rating

if it is explicitly stated or deduced from the reported information that the criterion is satisfied. If the criterion is not fulfilled or the information is missing, a score of 0 is given instead. For our systematic review, the scores were divided into three categories: Good quality = score ≥ 6, Fair quality = score of 4-5 and Poor quality = score ≤ 3 as in Van Criekinge et al. (39). The quality assessment on the PEDro-P scale was conducted by two reviewers independently (KP & SB). KP rated all the studies first and then SB assessed twenty percent of the total number of included studies to establish agreement between raters.

2.6 Meta-Analysis on Training & Transfer effects

Meta-analyses on the effects of WMU training on task performance in studies assessing task-based functional neuroimaging data were conducted using Review Manager 5.3 (RevMan) (40). The training group (TG) and Control group (CG) outcome scores, i.e. means and SDs, were extracted for both pre and post training brain imaging sessions. If there were multiple difficulty levels or conditions expressing the primary outcome, the average means and SDs were calculated. This is in accordance with the methodology from previous WM training meta-analyses (16, 17) where, in studies that used multiple tests to assess the same construct, the average of means and SDs was calculated to produce a single measure for each study. If the outcome scores were not reported in tables or in text, they were extracted from figures using the Plot Digitizer Software[2]. In cases where the standard error (SE) was given, it was converted to SD using the RevMan calculator. If the range was provided for individual studies instead of the SD value, then an SD estimate was calculated as the quarter of the range (41). If it was not possible to extract the SD from other data, then the average SD was calculated as an approximation for that study (42). If the study had more than one control group, they were combined into a single control group where the overall means and SDs were calculated based on the formulae provided by Higgins and Deeks (41). The difference between mean outcome score at pre and post training [Mean post - Mean pre] for each group was inserted into RevMan; a positive value suggesting performance was greater at post-test. The pooled

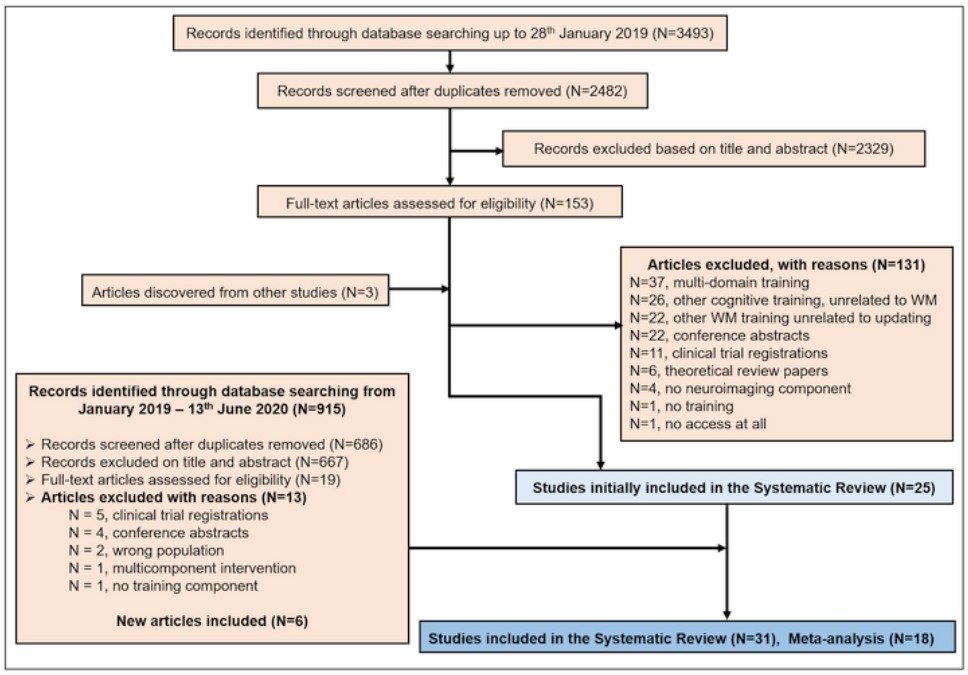
[2] http://plotdigitizer.sourceforge.net/

SD at pre-test was calculated and inserted for both TG and CG as recommended by Morris (3). This method has previously been used in other meta-analyses exploring the effects of cognitive training (17, 43). A random effects analysis model calculating the standardized mean difference (SMD) was selected in RevMan, to obtain SMD using Hedge's adjusted g (44) which is corrected for small sample bias. Consistent with Cohen's d (45), a Hedge's g was considered low at ≤0.20, moderate at ≥0.50, and large at ≥0.80. Heterogeneity was measured using the I^2 statistic and was considered low at 25%, moderate at 50% and large at 75% (46). Subgroup analyses based on the type of control group, training duration and type of transfer were conducted. Publication bias was examined using contour enhanced funnel plots created with the metafor package (47) within the RStudio environment (48) in R (49). An Egger's regression test (50) was conducted to examine funnel plot asymmetry.

3. Results

3.1 Study Selection

Of the 3493 records identified, 31 were included in this systematic review (Figure 1). Twenty-three of those were conducted in Europe (37, 51-72) four took place in Canada (73-76), three in the USA (77-79) and one in China (80). Eleven studies employed a randomized controlled trial methodology (37, 51-54, 59, 65, 71, 73, 74, 78), while eighteen used a quasi-experimental design (55-58, 60-63, 66-70, 72, 77, 79, 80) and two were case studies (75, 76). Twenty-seven of the studies included healthy adult participants, three included neurological populations (55, 75, 76) and one study included both (51).



[Figure 1 here]

3.2 Overview of healthy adult studies

The present review focuses primarily on studies with healthy adult samples, as this was the type of population investigated in most studies meeting the eligibility criteria. For this reason, information on the neurological samples is not presented in detail but summarized in section 3.10. The total number of healthy adult participants across studies was 955 (weighted mean age=31.94 (N=900), pooled SD=16.76 (N=900)). The total number of training group participants was N=464 (weighted mean age=34.02 (N=415), pooled SD=18.35 (N=415)), while those belonging to a control group were N=486 (weighted mean age=30.45, (N=448), pooled SD=15.21, (N=448)). If different studies shared the same sample, the dataset was only used once to calculate the total numbers of participants, means and SDs of age. Twenty-five of the healthy adult studies included a CG in their design (Table 1). The CG was either passive (37, 51-54, 56, 57, 61, 64, 65, 67, 68, 70, 80), active (58, 59, 71, 72, 77) or studies utilised both active and passive (66, 69, 73, 74, 78, 79) CGs. The remaining three studies had no CG (60, 62, 63). Participants trained for a total of 199.96 hours, ranging from 2.5 to 28 hours

(mean=9.52, SD=5.04) across studies. The total number of sessions varied between four and 55 (mean=16.67, SD=11.55), the training duration for each session ranged from 20 to 60 minutes per session (mean=38.93, SD=11.86) and total weeks of training ranged from one to 12 (mean= 4.29, SD=2.65). Further study details and information on the training protocols are summarized in Table 1 below.

[Table 1 here]

The neuroimaging measures used to evaluate the effect of WMU training are summarized in Table 2. Concentrating on the healthy adult studies, twenty-six used MRI (37, 51, 54, 56-74, 77-80) and only two used PET (52, 53). In eleven studies only the trained task, i.e., the criterion task, was performed in the scanner (51, 52, 58-60, 62, 63, 65, 72, 77, 79), while seven studies scanned both the criterion task and at least one untrained task, i.e., transfer task (37, 53, 61, 69, 71, 73, 78). In four studies, a transfer but not the criterion task was performed in the scanner (64, 66, 70, 80) and the remaining six studies did not assess task-based functional neuroimaging data (54, 56, 57, 67, 68, 74).

[Table 2 here]

3.3 Quality Assessment

Across studies on healthy adults, the PEDro-P score ranged from one to eight (Supplementary Table S1). Five earned a good rating, 13 were rated as fair and 10 as poor. Most of the studies failed to meet or report information concerning the following items: allocation concealment (item 3), blinding of subjects (item 5), blinding of assessors (item 7) and whether participants with available outcome measures received the treatment or control condition allocated (item 9). On the contrary, items 10 (between-group statistical comparisons reported for at least one key outcome) and 11 (both point measures and measures of variability provided for at least one key outcome) were most frequently met.

3.4 Training Effect: Healthy Adult Studies

The TG showed greater improvement, as assessed in terms of criterion task accuracy compared to the CGs, across all included studies irrespective of training protocol (Table 3). Reaction times also improved after training in the studies additionally reporting this outcome measure (54, 61, 65, 74, 79). For the studies employing criterion tasks with various difficulty levels, the training effect was greatest for higher levels of task difficulty. In addition, training duration as short as 2.5 (77) and 3 hours (65) produced a behavioural improvement.

[Table 3 here]

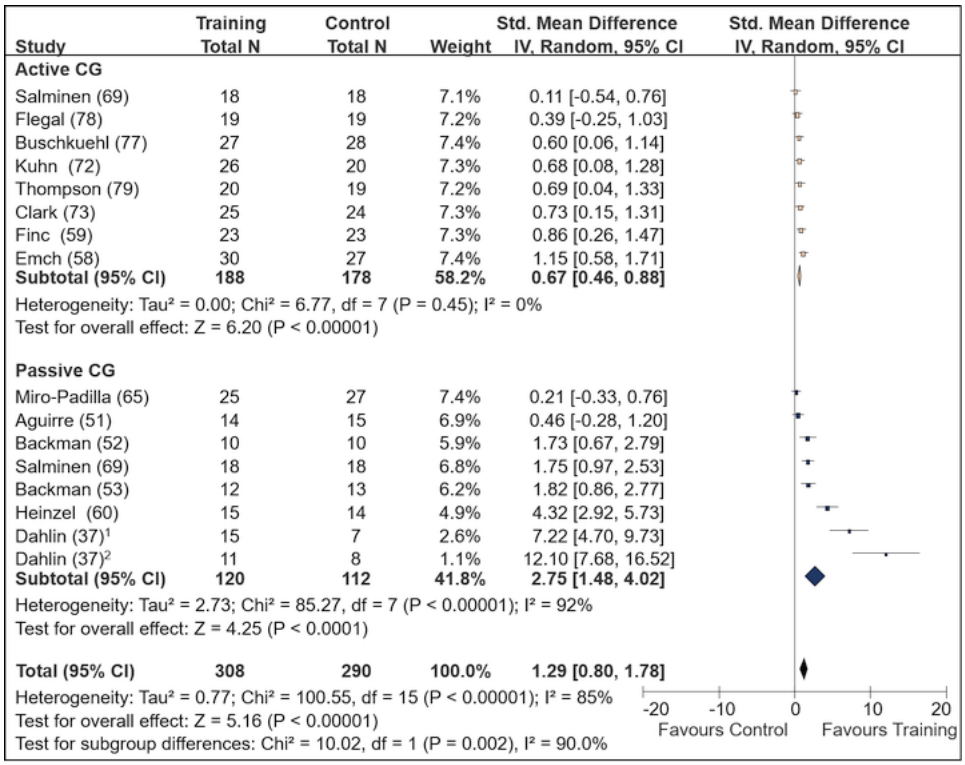
3.4.1 Meta-Analysis of Training effect in Healthy Adult Studies Assessing Task-Based Functional Neuroimaging data

Of the 22 healthy adult studies that assessed task-based functional neuroimaging data, 14 were included in a meta-analysis investigating training effects (Figure 2). One study was excluded as it did not report behavioural data on the scanned criterion task (71), three did not use a pretest-posttest control group design (60, 62, 63); and four assessed scanned transfer tasks exclusively without including a criterion task in their protocol (64, 66, 70, 80). Overall the training effect following WMU training was large, Hedge's g=1.29 (95% CI 0.80 to 1.78, Z=5.16, p<0.00001), with large heterogeneity across studies (I^2=85%). The training effect funnel plot exhibited signs of asymmetry indicating possible publication bias (Supplementary Figure S1); and the Egger's regression test yielded significant results (z = 9.36, p < .0001).

3.4.1.1 Control Group Sub-Group Analysis

Sub-group analyses were conducted to investigate whether heterogeneity across studies included in the meta-analysis was reduced by comparing the TG with the active control group (ACG) and passive control group (PCG) separately. The PCG sub-group analysis revealed a very large effect size of Hedge's g=2.75 (95% CI 1.48 to 4.02, Z=4.25, p < 0.0001). In contrast, the ACG sub-group analysis showed a moderate to large effect size of Hedge's g=0.67 (95% CI 0.46 to 0.88, Z=6.20, p<0.00001). Heterogeneity remained large for the PCG analysis (I^2=92%) while it reduced to zero for the ACG analysis (I^2=0%). There was also a significant

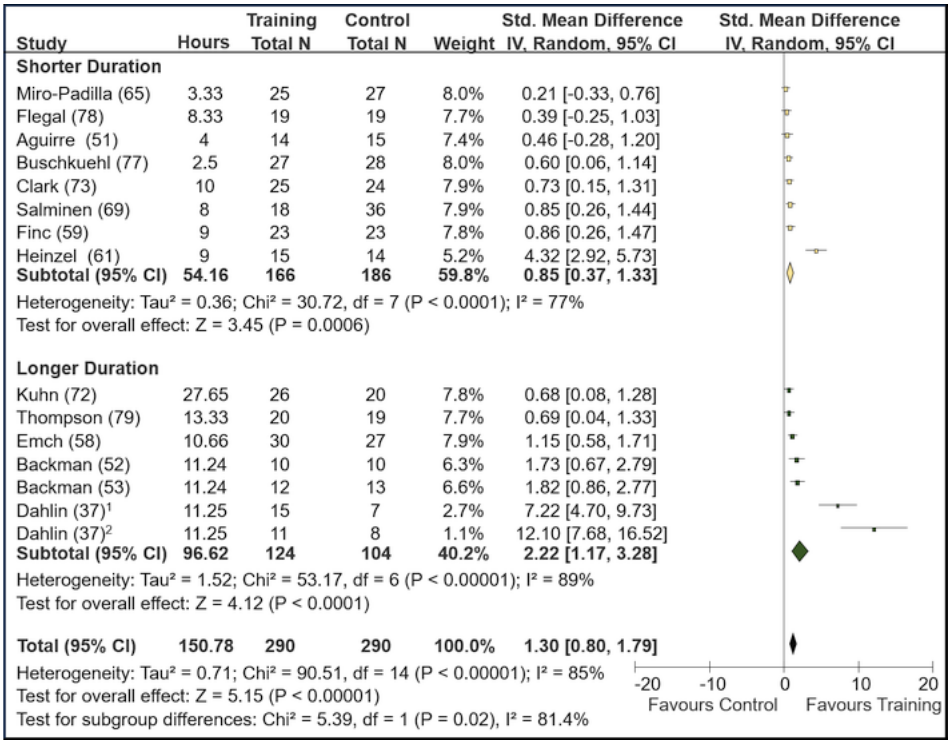
sub-group effect (χ^2=10.02, p=0.002) indicating that the type of control group significantly modifies the effect of training.



[Figure 2 here]

3.4.1.2 Training Duration Sub-Group Analysis

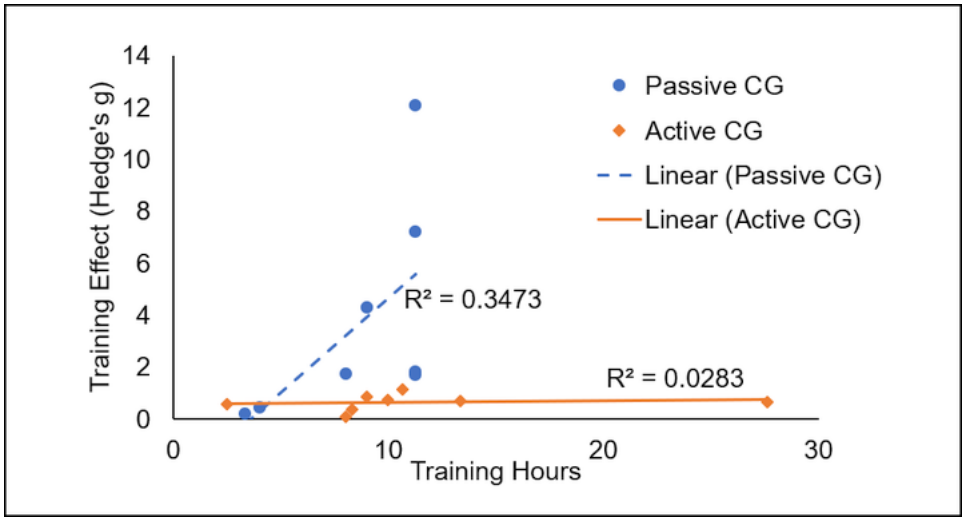
Further sub-group analyses were conducted to investigate if training duration impacted on the effect of WMU training. The median value for training hours across studies included in the meta-analysis was 10 (mean=10.05, SD=5.82) with those equal and below the median duration categorized as “shorter duration” and those above categorized as “longer duration”. Both subgroups exhibited large training effect sizes: shorter duration group Hedge’s g=0.85 (95% CI 0.37 to 1.33, Z=3.45, p=0.0006) and longer duration group Hedge’s g=2.22 (95% CI 1.17 to 3.28, Z=4.12, p<0.0001) (Figure 3). There was a significant subgroup effect (χ^2=5.39, p=0.02), indicating that training duration significantly modified the effect of training, favouring training of longer duration. However, due to large heterogeneity within each group (shorter duration sub-group I^2=77%; longer duration sub-group I^2=89%), the overall effect sizes should be interpreted with caution.



[Figure 3 here]

3.4.1.3 Relationship between Control Group & Training Duration

We further plotted training duration against the effect of training for ACG and PCG sub-groups analyses (Figure 4). The training effect size for studies comparing the TG against ACG remains stable regardless of training duration while a linear upward trend is apparent in the training effect size for studies comparing the TG against PCG as the hours of training increase.



[Figure 4 here]

3.5 Transfer effect: Healthy Adult Studies

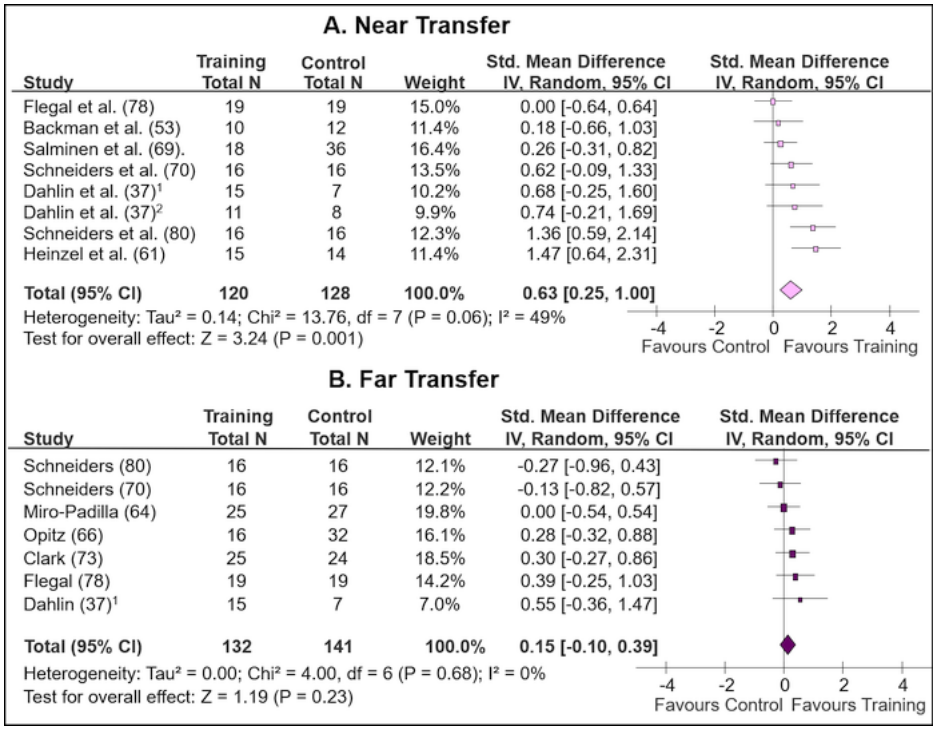
Of the 13 included studies assessing near transfer effects following WMU training (37, 52-54, 58, 61, 62, 69, 70, 72, 77, 78, 80), mixed results were reported, (Table 3). The studies by Backman et al. (53); Biel et al. (54); Flegal et al. (78); Kuhn et al. (72) and the older adult training group in the study by Dahlin (37); did not find significant near transfer effects to untrained tasks in the same cognitive domain. On the contrary, the studies by Backman et al. (52); Buschkuehl et al. (77); Emch et al. (58); Heinzel et al. (61); Schneiders et al. (70, 80); and the young adult training group in the study by Dahlin et al. (37); all found evidence of a near transfer effect after WMU training. Finally, Heinzel et al. (62) and Salminen et al. (69) used single and dual versions of a delayed match to sample task and a WMU task, respectively, to assess near transfer. Both studies found significant effects only for the dual versions of the task and no effects for the single versions.

Far transfer following WMU training was assessed in nine of the included studies (37, 54, 64, 66, 70, 71, 73, 78, 80), (Table 3). Biel et al. (54); Schneiders et al. (70, 80); Miro-Padilla et al. (64) and the young adult training group in the study by Dahlin et al. (37) did not find significant far transfer effects following WMU training. On the contrary, Flegal et al. (78) found evidence of far transfer for the highest difficulty level of an untrained episodic memory task; Opitz et al. (66) reported improved performance in an untrained Chinese orthographic task assessing far transfer, while Schweizer et al. (71) reported greater reduction in emotional distress exhibited by the TG compared to the CG following emotional WMU training. Clark et al. (73) utilized two tasks to assess far transfer, the Raven's Standard Progressive Matrices (RSPM) and a lexical decision task and reported better performance for the TG compared to the CG on both tasks. However, the authors further explained that the RSPM task effect was driven by worse post-training performance in the CG compared to the TG, while significant differences between groups at baseline accounted for the far transfer effect for the lexical decision task.

3.5.1 Meta-Analysis of Transfer Effects in Healthy Adult Studies Assessing Task-Based Functional Neuroimaging data

Of the 22 healthy adult studies assessing task-based functional neuroimaging data a total of ten included transfer tasks in their protocol; with three investigating near transfer effects exclusively (53, 61, 69), three assessing a far transfer task only (64, 66, 73) and four examining both near and far transfer tasks (37, 70, 78, 80).

The near transfer effect after WMU training was moderate, Hedge's g=0.63 (95% CI 0.25 to 1.00, Z=3.24, p=0.001) with moderate heterogeneity across studies (I^2=49%), (Figure 5A). On the contrary, the analysis of far transfer exhibited a small non-significant effect, Hedge's g=0.15 (95% CI -0.10 to 0.39, Z=1.19, p=0.23) and zero heterogeneity across studies (I^2=0%), (Figure 5B). The Egger's test for funnel plot asymmetry yielded non-significant results for both near (z = 1.30, p = 0.19) and far transfer (z = 0.26, p = 0.79) (for further details please see Supplementary Figures S2 and S3).



[Figure 5 here]

3.6 Training task: Functional Activity Changes in Healthy Adult Studies

Most of the reviewed fMRI studies found decreases in BOLD activity during the criterion task performance after WMU training (51, 58, 60, 61, 65, 71, 73, 75, 78, 79), (Table 4). Despite varying in terms of training protocol, task type and modality, overall these studies showed a similar pattern of results: decreases were detected primarily in: 1. frontal areas, i.e., frontal pole, superior frontal gyrus, DLPFC, the pre-motor and insular cortex, the cingulate gyrus, and 2. parietal areas, i.e., intraparietal sulcus, inferior parietal lobule. An exception to this pattern was increased BOLD activity in fronto-parietal areas and striatum reported for the older adult training group in the study by Dahlin et al. (37).

Backman et al. conducted two similar PET studies (52, 53) and found decreases in raclopride binding to D2 receptors in the striatum, translating to increased dopamine (DA) release as a result of WMU training. Previous research has revealed a link between BOLD activity and DA release measures (81), and thus an increase in DA release is linked with an increase in striatal BOLD activity. Bushchkuehl et al. (77) conducted an ASL study and also found increases in signal magnitude indicative of increased perfusion, a surrogate for functional activity, on the criterion task in frontal and occipital areas after only 2.5 hours of training.

Buschkuehl et al. (77) additionally reported both increases and decreases in perfusion at rest. Increases were evident in the left precentral gyrus and left parietal angular gyrus while a decrease was found in the right postcentral gyrus.

Salminen et al. (69) found BOLD decreases in fronto-parietal regions, and an increase in the pre-central gyrus, on the criterion task after WMU training. For the young adult training group, Dahlin et al. (37) reported decreases in fronto-parietal areas and increases in the striatum, temporal and occipital regions.

Studies employing more than two scanning sessions provide valuable insight into the dynamics of training-related activation increases and decreases elapsing over time. Hempel et al. (63) and Kuhn et al. (72) reported initial BOLD increases between sessions 1 and 2, i.e.

pre training and early training fMRI session respectively, followed by decreases between sessions 2 and 3, i.e. from early training to post-training. More specifically, Kuhn et al. (72) reported striatum increases at first followed by striatal and frontal decreases after several dozen intervening sessions of training, while Hempel et al. (63) reported an initial BOLD increase at the right intraparietal sulcus and superior parietal lobe two weeks into a four-week training regimen, and a subsequent decrease in these areas post-training.

3.7 Transfer Task: Functional Activity Changes in Healthy Adult Studies

3.7.1 Near Transfer

Dahlin et al. (37) found post-training BOLD increases in striatum and frontal, parietal and temporal cortex when assessing a near transfer task in a young adult training group, while no significant changes were reported in an older adult training group. Salminen et al. (69) found increased BOLD activity in the striatum, cuneus and calcarine gyrus for a near transfer task. Schneiders et al (70, 80) reported decreases in BOLD activity as a result of n-back training in two different studies. The first involved decreases in the middle frontal gyrus for a visual n-back near transfer task (70), and the second found decreases in the IFG for an auditory n-back near transfer task (80). Heinzel (61) reported BOLD activity decreases in middle and superior frontal areas specifically for the combined 3&5 update condition of a near transfer task, in a study with older adults. The study by Flegal et al. (78) interrogated a priori subcortical ROIs that revealed no significant differences in BOLD activity changes between the TG and ACG.

Finally, in a PET study, Backman et al. (53) found increased striatal DA release, linked with an increase in striatal BOLD activity as explained above, for an n-back near transfer task.

3.7.2 Far Transfer

Clark et al. (73) found increased activity post-training in frontal regions as well as the precentral and postcentral gyrus for the highest level of difficulty in a far transfer task. Schweizer et al. (71) reported increased BOLD activity in the superior temporal gyrus associated with the emotional regulate condition in a far transfer task. On the other hand, Miro-Padilla et al. (64) reported activity decreases in the right DLPFC for a far transfer auditory

attention task after 3.33 hours of training. Opitz et al. (66) found decreased BOLD activity in the fusiform gyrus for an untrained Chinese orthographic task, only for the PCG, while no changes were reported for the TG or ACG. Lastly, Dahlin et al. (37), Schneiders et al. (70, 80) and Flegal et al. (78) did not report any significant BOLD changes when assessing far transfer tasks after WMU training.

[Table 4 here]

3.8 Functional Connectivity Changes: Healthy Adult Studies

Only a handful of studies explored changes in functional connectivity as a result of WMU training (Supplementary Table S2). Thompson et al. (79) observed an increase in functional connectivity for all pairings of prefrontal and parietal ROIs, including lateral prefrontal and parietal cortex, for the 2-back load condition of the criterion task, whereas Heinzel et al. (60) did not find any significant connectivity changes in the WM network as a result of training. Assessing training-induced changes in functional brain network modularity across four scanning sessions, Finc et al. (59) reported increased recruitment of the fronto-parietal and default mode systems for the TG post-training, while the integration between these two systems decreased post-training. Integration changes between the subcortical and other systems was also explored with decreases reported at the early stages of training and increases post-training between the subcortical and default mode systems. The exact opposite pattern was revealed for the integration between the subcortical and dorsal attention, ventral attention, cingulo-opercular and auditory systems, in that increases were reported at first and decreases at the end of training.

3.9 Structural Changes: Healthy Adult Studies

The pattern of results regarding training-induced changes on structural imaging measures was not straightforward with most studies reporting null findings (Supplementary Table S3). The studies by Heinzel et al. (60) and Biel et al. (54) did not find significant GM volume changes, myelination or iron levels (54). Likewise, Lawlor-Savage et al. (74) reported no changes in cortical surface, thickness or volume after training. Colom et al. found volume preservation for

the TG (56) and in the context of decreased gray matter volume for the CG in bilateral temporal lobe (57). When carrying out further analyses on the same dataset as in Colom et al. (56, 57), Roman et al. (67) reported mean cortical thickness changes in the right ventral frontal and right middle temporal cortex, revealing minor thickening for the TG and minor thinning for the CG. They also found cortical surface area changes in the right pars opercularis and right posterolateral temporal cortex, revealing a small expanding effect for the TG and a small contracting effect for the CG. Finally, Roman et al. (68) conducted network- based statistics in the same dataset as in Colom et al. (56, 57) and Roman et al. (67) and identified a sub-network including frontal, parietal, temporal, subcortical regions and the insula where changes after training were more pronounced for the TG. The left middle temporal region was identified as the most highly interconnected area with connections to the bilateral basal forebrain, left parahippocampal area, left pallidum, left supramarginal and left parietal area, right insula, right accumbens, right postcentral gyrus, right pars opercularis and right pars triangularis. There was increase in structural connectivity for the TG post training in this network while no changes were observed for the CG. Furthermore, the authors reported increases in the connectome topological properties of global efficiency and strength in this sub-network for the TG while no changes were observed for the CG.

3.10 Neurological Populations: An overview of findings

Four studies included in this review assessed neurological samples; two of those were stroke case studies conducted in Canada (75, 76), and the other two took place in Europe and included adults diagnosed with multiple sclerosis (51, 55), (Table 1). Only one of the studies employed a pretest-posttest control group design (51) while the rest did not include a CG (55, 75, 76). All studies applied an n-back training protocol and the training duration ranged between four and 20 hours. All studies included an fMRI task-based analysis while none explored changes in the brain's functional connectivity or structure changes following WMU training, (Table 2).

Participants improved their criterion task accuracy as a result of WMU training across studies (51, 75, 76), (Table 3). Aguirre et al. (51) did not report data for the healthy controls (HC) and multiple sclerosis (MS) participants separately; thus the exact training effect for each population could not be analysed. Bonzano et al. (55) did not assess performance on the criterion task but examined transfer effects for tasks performed inside and outside the scanner. Improved performance was found on all tasks of the Rao's Brief Repeatable Battery of Neuropsychological Tests (BRB-NT) post-training compared to pre-training, although it is important to note that this study did not include a CG.

As with the training-related behavioural data, Aguirre et al. (51) did not report neural changes following WMU training for the different participant groups separately; nevertheless, fronto-parietal activity decreases were found for both HC and MS, (Table 4). Furthermore, Leung et al. (75, 76) reported a mixture of BOLD increases and decreases in fronto-parietal and temporal areas after training. Finally, Bonzano et al. (55) assessed fMRI performance on a far transfer task exclusively and reported decreases in fronto-parietal areas post-training compared to pre-training.

4. Discussion

This is the first systematic review assessing cognitive and neural outcomes following training of the WMU process specifically. We concentrated on neuroimaging studies in adults and further conducted meta-analyses to investigate the effect of training, and transfer to untrained tasks, in studies assessing task-based functional neuroimaging data. Cognitive outcomes across the included studies reveal a clear pattern consistent with previous meta-analyses in the wider field of WM training. The neural changes after WMU training were assessed qualitatively and examined for both training and transfer tasks. These data reveal interesting training-related patterns with greater consistency in fronto-parietal cortical regions than subcortical areas. We interpret our results in relation to previous theoretical models.

4.1 Training Effect: Healthy Adult Studies

A meta-analysis of published studies indicates that WMU training can significantly improve cognitive performance in adults. However, the funnel plot for the training effect exhibited significant asymmetry indicative of publication bias. The observed large overall training effect in the reviewed data could be overestimated and biased from studies with small sample sizes, considerable variability and large effect sizes. When conducting sub-group analyses according to the type of control group, the training effect size was very large for studies with a passive control group, while a moderate effect was revealed for studies with an active control group. There was a significant difference between the training effect sizes from the control group sub-group analyses. At the same time, the large heterogeneity value in the PCG comparison in contrast to no heterogeneity for the ACG comparison suggests that studies employing a PCG introduce greater heterogeneity or noise in the data which could be possibly overestimating the training effect sizes. Similar findings have been reported in previous meta-analyses examining the influence of type of control group on transfer effects (15, 17). PCG designs do not control for a potential placebo effect thus making it difficult to discern whether the effect sizes stem from true training gains or perhaps mediated by non-specific factors such as increased effort (15). On the other hand, employing an ACG in which participants practice an alternative but similarly challenging task bears the risk of underestimating the effects of training (23). For this reason, there should be a dynamic balance between a no contact control group and a cognitively challenging control group such as employing a lower level non-adaptive task paradigm (23).

We further inspected how training duration affects the WMU training effect and found some evidence for an association between training duration and training effect size, although heterogeneity within both shorter and longer duration sub-groups was large. Finally, there seems to be a linear upward trend for the training effect size as the total hours of training increase for studies with passive control groups, while the effect size is insensitive to training duration for studies with active control groups.

4.2 Training Task: Functional Activity Changes in Healthy Adult Studies

The most consistent pattern of training-related changes involved BOLD activity decreases in fronto-parietal regions. These include frontal areas such as the frontal pole, superior frontal gyrus, DLPFC, pre-motor and insular cortex, cingulate gyrus, and parietal areas such as the intraparietal sulcus, inferior parietal lobule. The locations are consistent with a WM fronto-parietal network already established in the neuroimaging literature (5, 7, 12). Decreases in functional activation are thought to reflect neural efficiency, i.e. fewer resources needed to perform the same task after training than before training (24). This interpretation is consistent with the concept of plasticity proposed by Lövdén et al. (2) in which the neural system responds to a prolonged situation of environmental ″demands″ (e.g., a continuously challenging cognitive task) exceeding functional ″supply″ (i.e., neural resources) with plastic changes.

Increases in functional activation after WMU training were observed in an older adult group in fronto-parietal regions and striatal areas (37). This is in direct contrast to other studies which also included older adult training groups but reported decreases in fronto-parietal activity instead (58, 61); a neural response pattern similar to that seen in young adults. Previous literature suggests that older adults often exhibit greater activation compared to young adults (82-84) and one explanation for this is a compensatory use of neural circuits, known as the CRUNCH model (85). This model posits that older adults reach a peak in functional activity at lower difficulty levels than young adults, indicating that the point at which neural resources reach maximum capacity differs with age. Iordan et al. (86) tested the CRUNCH hypothesis model on a within-subject intervention design with young adult and older adult groups and confirmed that, irrespective of age, WM training leads to functional activity decreases (i.e. fewer resources needed to perform the task after training), consistent with the studies by Heinzel et al. (61) and Emch et al. (58). The results further suggest a shift in the peak activation as a result of training, i.e. neural resources reach maximum capacity at higher difficulty levels than before the intervention. However, the older adult training group in the study by Dahlin et

al. (37) was not found to exhibit overactivation compared to the young adult training group, and its reported increase in striatal activation resulted from significant post-training activation that was not present at the pre-test session. Additionally, the older adults' behavioural performance was quite poor at pre-training. These findings suggest the anomalous result of increased fronto-parietal activity post-training observed by Dahlin et al. (37) could be explained by the older adult group experiencing the criterion task as markedly more difficult than the young adult group pre-training, for which a post-training shift in the peak activation via training-induced plasticity (86) would in fact produce relative increases in activity.

A mixture of activity increases and decreases over time were reported in studies that employed three scanning sessions, i.e. pre-training, early training and post-training. Initial striatal increases followed by striatal and frontal lobe decreases after training were reported by Kuhn et al. (72), while Hempel et al. (63) reported an initial BOLD increase and subsequent decrease at the right intraparietal sulcus and superior parietal lobe. Buschkuehl et al. (77) also reported increases in ASL perfusion, a surrogate of BOLD activity, in superior frontal and postcentral gyrus together with superior and middle occipital gyrus after a brief 2.5 hours of WMU training. Thus, it should be borne in mind that some variability in the direction of activation changes across the other reviewed studies could be due to a dynamic process being captured at a single post-training timepoint for comparison to a pre-training baseline, defining an interval that ranges widely across studies.

Doyon and Benali (87) proposed a fast-early and a slow-late stage model of motor learning in which the cortico-striatal and cortico-cerebellar systems contribute differentially to the learning process, where activity changes in the two systems are observed at different learning stages. Lustig et al. (88) hypothesized that if this motor learning model is applied in cognitive training, then fronto-parietal increases should be observed at the beginning, followed by potential decreases or a mixture of increases and decreases in these networks. For WMU training, studies by Hempel et al. (63) and Kuhn et al. (72) support this hypothesis of early stage activity

increases and late stage decreases. The ASL study (77) further corroborates this model with evidence of increased perfusion after only 2.5 hours of training.

Patterns of activation changes following WMU training appear less clear in subcortical regions. The two PET studies by Backman et al. (52, 53) reported increased dopamine release specifically involving the striatal region which is consistent with training-induced functional activity increases in the striatum. Even though a link between DA release and BOLD activity has been previously established (81), this pattern of results should be interpreted with caution due to the different measures employed by the PET and fMRI methodologies, i.e. altered neurotransmitter synthesis and BOLD activation changes respectively. Using fMRI, Flegal et al. (78) and Kuhn et al. (72) reported striatal activity decreases after WMU training, while Dahlin et al. (37) found a striatal increase for a young adult training group. A commonality in these studies setting them apart from others that did not report subcortical activation changes is that all used memory updating task paradigms, rather than an n-back training task in which WM load varies along with WMU demand (perhaps accounting for the predominance of activity changes within the WM fronto-parietal network in studies that used n-back training tasks). One reason why the direction of striatal activity change after training is inconsistent across studies could be that decreases were observed for training groups compared to an active control group (72, 78), while increases were observed in a passive control group comparison (37) .

Our findings are consistent with those from reviews of the wider WM training literature in that the neural pattern of activation changes exhibited decreases, increases and mixture of decreases and increases post-training. A summary of these changes after WMU training suggests the following: 1. Robust evidence of BOLD decreases in fronto-parietal regions across studies, 2. Dynamics of activity changes differ at the fast-early and slow-late learning stages, showing an initial increase and a subsequent decrease in BOLD activity, 3. Training-related striatal activation changes are found when a memory updating task is employed rather than an n-back task; with some studies reporting increases and some reporting decreases.

Nyberg and Eriksson (89) proposed a subcortical dopaminergic updating system in which dopaminergic neurotransmission and striato-cortical interactions are involved in WMU and the striatum constitutes a major subcortical node for updating. Dopaminergic neurotransmission is also central to a model developed by Cools and D'Esposito (90) which views cognitive control as a multifactorial phenomenon where a dynamic equilibrium between cognitive stability (manifested in the prefrontal cortex) and flexibility (manifested in the striatum) is essential. This model relies on the qualitatively different functional DA roles in the PFC and striatum. Recent findings propose that striatal DA plays a role in WM and cognitive control by serving as the gate mechanism crucial for flexibly updating the current goal representations in the PFC, while the PFC DA enhances stability of these representations by strengthening distractor resistance and attenuating the PFC networks (90). The authors hypothesize that our behaviour needs to flexibly update according to relevant changes, e.g. switching between different tasks, but also remain stable when these are irrelevant, e.g. focussing on a task without getting distracted by external factors. Flexibility and stability are ascribed as two functionally distinct and opposing mechanisms that ultimately work together, complement each other and are manifested in the striatum and PFC respectively (90).

We therefore suggest the striatum responds differentially to learning and/or cognitive training compared to the fronto-parietal network and that makes it a key factor to explain the pattern of results reported above. We propose that the hypothesized PFC involvement in cognitive stability is supported by the consistency in fronto-parietal BOLD decreases after WMU training across studies included in this review. Previous reviews have reported activity changes in fronto-parietal and subcortical areas after WM training (13, 14, 19, 20) but the present review is the first to focus solely on the WMU process, finding a consistent pattern of decreased fronto-parietal activity post-training. In contrast, we view the inconsistencies in the striatal activity changes as a manifestation of cognitive flexibility. Based on the theoretical framework of adult cognitive plasticity by Lövdén et al. (2) combined with models of the striatum as a major node for updating (89), we suggest that neural changes in the striatal region are a

manifestation of Lövdén's concept of flexibility, i.e. the neural system's existing ability to adapt effectively to environmental demands and utilise the necessary neural processes for performing a given task. Our analyses suggest that significant changes in striatal activity are found only after training on studies employing a memory updating task paradigm. Even though both memory updating and n-back task paradigms tap into the WMU process, they also entail distinct cognitive processes. Memory updating tasks involve storage and updating, e.g. the WM load remains stable even though the updating demands vary across task difficulty levels. N-back tasks involve simultaneous storage, monitoring, maintenance and updating, e.g. the WM load also changes as a result of varying the updating demand. For this reason, we suggest that the memory updating paradigms are more likely to specifically target the WMU process and we will refer to them as "highly targeted" memory updating tasks, e.g. matrix updating or numerical memory updating. We further propose that these highly targeted WMU tasks can successfully "trigger" the neural system's flexibility which is manifested in the striatal changes after training. However, this is a speculative explanation of our findings and should be interpreted with caution due to the small sample of reviewed studies.

4.3 Transfer Effect: Healthy Adult Studies

For transfer of training gains to untrained tasks, WMU training was found to improve performance on near transfer tasks (same cognitive domain) but not far transfer tasks (different cognitive domain). Again, our findings are consistent with previous syntheses of cognitive outcomes from WM training (17, 18) in reporting a medium-sized near transfer effect and a non-significant far transfer effect. Our transfer results seem consistent with the notion that overlapping cognitive processes are necessary for transfer to occur as previously suggested (20) and that would theoretically explain the lack of far transfer, i.e. when the criterion and transfer task do not share the underlying process of WMU.

However, it is important to point out there are discrepancies in what authors identify as near and far transfer across studies. These terms are not used consistently in the cognitive training literature, contributing to the difficulty of defining the concept of transfer adequately and

ultimately reaching a consensus. In our review of WMU training studies, we categorised transfer tasks as near or far by following the authors' own classifications and we further collapsed across task difficulty levels and averaged performance across multiple tasks to minimise bias in our meta-analysis to the greatest extent possible. However, we acknowledge the complexity of this issue and would like to draw attention to the fact that our reported findings regarding transfer effects ultimately rely heavily upon the definitions of near and far transfer within each reviewed study.

Moreover, there were not enough reviewed studies with transfer task data to allow sub-group assessment for type of control group and therefore we are unable to make claims regarding the influence of active and passive control groups on WMU training interventions. We suggest that including sub-analyses to investigate the training duration, the training paradigm, and control group can potentially clarify the issue of near and far transfer further.

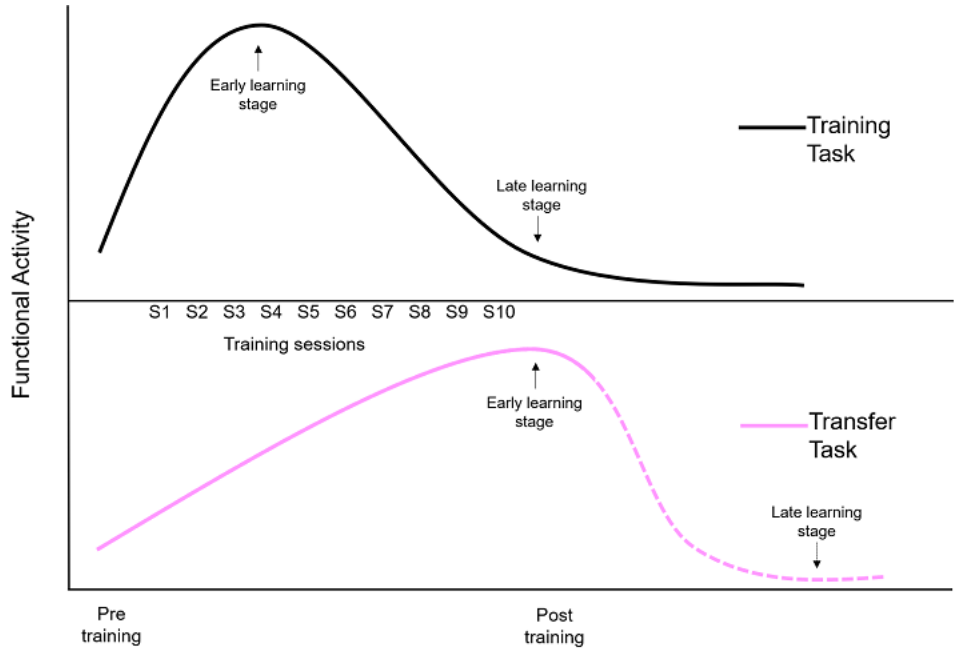
4.4 Transfer Task: Functional Activity Changes in Healthy Adult Studies

Most studies and previous literature reviews of WMU training have not focused on training-related neural changes on transfer tasks. We found activity increases and decreases, primarily in frontal and striatal regions, for scanned transfer tasks after WMU training. Overall, studies reported functional activation increases (37, 52, 69, 71, 73) consistent with the WM training meta-analysis by Salmi et al. (12) reporting IFG and striatum increases in transfer tasks. On the contrary, other studies observed no significant changes in activity (37, 66, 78). A few studies reported transfer task activity decreases after WMU training, and a closer look reveals they are distinct from the rest. The study by Heinzel et al. (61) involves older adults whose neural response is different compared to young adults (82-84). Even though the study by Miro-Padilla et al. (64) exhibited decreases in a far transfer task following n-back training, there were no significant behavioural transfer effects and thus we are unable to assign a meaningful interpretation to these neural findings. The remaining two studies by Schneiders et al. (70, 80) differ in their categorization of transfer tasks; they use the definitions of intra-modal and across-modal general control task instead of near and far transfer task, respectively. We

suggest that what the authors view as intra-modal transfer (performance on a visual 2-back task with novel stimuli following training with a visual adaptive n-back task) is what many cognitive training researchers would consider a measure of the criterion i.e. trained criterion task; while what they authors view as across-modal transfer (performance on a visual 2-back task following training with an auditory adaptive n-back task) is closer to a typical measure of near transfer. Following that logic, then frontal BOLD decreases for the intra-modal tasks are consistent with the fronto-parietal reductions for the training tasks in other reviewed studies, while the lack of activity changes for the across-modal tasks suggest no neural changes taking place for a near transfer task after WMU training.

Returning to the fast-early and slow-late stage model first applied to motor learning (87), we propose this can be extended to account for the commonly observed activation increases for transfer tasks following WMU training. Similar to the dynamic activation increases and decreases elapsing over time for training tasks scanned early in training and then again later in training (63, 72), we suggest that activation profiles for transfer tasks also follow the same inverted U-shape pattern, but at a different rate reflecting their less frequent exposure to training study participants. Due to this, there is a hypothesized time-lag in the activation curve as a function of time for transfer tasks, compared to that of the training task. The post-training activity increases frequently reported for transfer tasks result from training on the criterion task, and although its post-training activation changes on the criterion task are most frequently reported as decreases, both profiles can be represented by the same schematic model of training-related neural changes (Figure 6). Repeated exposure to, and practice with, the training task is associated with functional changes observed as early-stage activity increases (on the scanned criterion task) followed by late-stage activity decreases that may represent neural efficiency resulting from plastic changes induced by WMU training. The most common experimental design for cognitive training studies assessing task-based functional neuroimaging data is to scan transfer tasks at one post-training session, and although participants have had repeated exposure to the training task at this point the post-training

transfer task is still relatively novel and challenging, thus performance is still effortful—similar to a criterion task at the early stage of learning—and the activation change from baseline is observed as an increase. The dashed line following the post-training scanning session for the transfer task in Figure 6 represents a predicted functional activity decrease that would eventually occur if participants were repeatedly exposed to the transfer task, thereafter, consequently approaching the slow-late learning stage.



[Figure 6 here]

4.5 Other Neural Changes

Only three of the reviewed studies examined functional connectivity changes following WMU training, restricting the possibility of drawing definitive conclusions. Thompson et al. (79) reported connectivity increases within fronto-parietal ROIs for the training group, consistent with previous WM training literature (25, 26). Finc et al. (59) was the only study to conduct an extensive analysis on training-related functional connectivity modulations on large scale brain networks. Increased fronto-parietal and default mode system recruitment was reported post-training, while the integration between these two systems exhibited decreases post-training. Another interesting finding was a dynamic modulation of the integration between the subcortical and other systems. Decreases between the subcortical and default mode systems

were reported at the early stages of training and increases post-training, while the exact opposite pattern was revealed for the integration between the subcortical and dorsal attention, ventral attention, cingulo-opercular and auditory systems, in that increases were reported at first and decreases at the end of training. Heinzel et al. (60), on the other hand, did not find significant functional connectivity changes post-training in the training group for any of the difficulty levels, however that null effect could be due to the lack of a training vs control group comparison.

We cannot draw conclusions on the structural changes taking place after WMU training, as from the seven relevant studies in this review, four constitute different analyses of the same dataset while the other three found no significant training-related changes in gray matter volume (54, 60, 74), surface and thickness (74). The studies by Colom et al. (56, 57) and Roman et al. (68) reveal an inconsistent pattern of gray matter changes where volume preservation in the training group was reported in bilateral temporal lobe in one study (56) and an increase in volume in the right temporal lobe, left posterior cingulate cortex and right cerebellum in the other (57). The only study examining structural connectivity reported an increase in a fronto-parietal network after WMU training (67), consistent with an earlier WM training study (32).

4.6 Neurological Populations

Only a handful of the reviewed studies included neurological samples (51, 55, 75, 76), thus making it difficult to draw solid conclusions. However, these studies provide promising results suggesting that adults who have sustained damage to the brain also seem to benefit from a WMU intervention and improve their cognitive performance on the criterion task. Furthermore, they exhibit training-related fronto-parietal decreases similar to those reported in healthy adult studies. Nevertheless, it is evident there is a need for additional neuroimaging studies with a pretest-posttest control group design examining the effects of WMU training in neurological disorders. The application of research findings in a clinical setting depends upon researchers

designing and validating cognitive interventions with the objective to provide optimized and evidence-based training regimes for populations with cognitive impairments.

4.7 Summary

WMU training can significantly improve cognitive outcomes and produce moderate near transfer effects while there is currently no evidence for far transfer effects, consistent with previous reviews on WM training. The data included in this systematic review are indicative of publication bias, suggesting that studies with smaller samples exhibiting large training effects were more likely to have been published, which could potentially overestimate the overall effect size. Furthermore, WMU training effect sizes are significantly larger in studies comparing the training group to a passive control group than to an active control group. When comparing shorter and longer training durations, there was a significant sub-group effect suggesting that longer duration produces a larger training effect, as suggested by Von Bastian & Oberauer (23). However, our results indicate that this is true only for passive control group comparisons, while the training effect size in active control group comparisons remains unchanged as the training hours increase.

Our review reveals a fairly homogeneous pattern in neural outcomes regarding the training-related changes in functional activity. We hypothesized that the consistency in fronto-parietal activity decreases are a sign of the prefrontal cognitive stability while the discrepancy in striatal changes is an indication of cognitive flexibility. We further propose that employing a highly targeted WMU task training protocol with adaptive difficulty can successfully trigger training-related changes in the brain's system, which is an indication of plasticity. Our results also support a fast-early and slow-late stage model of learning in cognitive training, following an initial increase and a subsequent decrease in fronto-parietal activity as hypothesized by Lustig et al. (88). We further applied this learning model to explain the functional activity increases exhibited for transfer tasks post-training, suggesting the transfer activation profile is similar to that of the training task but slower, i.e. the response is lagged. This is the first review reporting

consistent neural patterns of activation post-training and we attribute this to our inclusion of studies training the updating process of WM specifically.

4.8 Limitations

The reviewed studies are not standard randomized clinical trials, rather the majority are quasi-experimental cognitive training neuroimaging studies. Nevertheless, such experimental designs are standard practice in human neuroimaging research due to practical limitations involving costs, limited personnel and time constraints. Consequently, the methodological quality of the included studies based on the PEDro-P Scale was generally modest and thus the results should be interpreted with caution. At the same time, the neuroimaging methodological quality could not be similarly assessed due to the lack of a standard quality scale comparable to the PEDro-P.

Overall there was a small number of included studies, due to our specific focus on neuroimaging studies with a pretest-posttest design targeting the WMU process exclusively in order to limit heterogeneity across studies. For the same reason, the small number of reviewed studies with transfer task data precluded a control group sub-group analysis on the transfer effect sizes. This would have the potential to reveal a significant difference between the active and passive control sub-groups and therefore clarify the mediators of far transfer. Similarly, our proposed interpretation of the functional activity changes in the transfer tasks following WMU training relies on a small number of studies and thus should only be considered speculative at this point and in need of testing with additional data. For the same reason, specific conclusions for studies assessing functional connectivity and structural imaging changes after WMU training could not be drawn.

There was also an overall lack of assessment on measures of everyday function in the reviewed studies and therefore we cannot be certain of the WMU training impact on daily living. Finally, the limited number of studies involving neurological populations makes it difficult to draw conclusions on WMU training efficacy in adults with brain damage or the impact on their ability to improve everyday functioning.

4.9 Conclusions

We conclude that WMU training can successfully promote plasticity under Lövdén's theoretical framework (2) as exhibited by improved cognitive performance, near transfer of training gains and indirect alterations in the structure of the brain's system evidenced by fronto-parietal and striatal functional activity changes post-training. Neural changes associated with WMU training follow a fronto-parietal fast-early activity increase and a late-slow decrease, while those associated with transfer of training appear to follow the same pattern albeit with a lag. A cognitive training protocol targeting the WMU process specifically can successfully trigger the neural system's flexibility manifested by the involvement of the striatum which is considered a major subcortical node for updating. Cognitive training studies are recommended to compare the training intervention against active control groups and employ a highly targeted WMU training protocol.

Future studies should additionally examine changes in measures of the brain's functional connectivity and structure as well as include a third scanning point when possible to improve our understanding of the neural mechanisms behind plasticity as well as the dynamic patterns of learning. Even though adding a third time-point in a longitudinal neuroimaging study can be quite challenging in terms of resources needed, evidence shows this can shed light into the dynamic patterns of neural modulation at different stages of training. There is no single right answer to the question of when the additional time point should be placed, as this is directly related to the specific research question the researcher wishes to pose. For example, in order to explore the plausibility of predicted functional activity increases early in training followed by decreases at later stages, then one would theoretically add a scanning session very early in the training period, e.g. after only a few hours of training. On the contrary, to examine whether the activation profiles for transfer tasks follow the same hypothesized inverted U-shape pattern as the training task, then the additional time-point would need to be placed after the end of the training period.

Finally, even though our interests include the cognitive and neural effects of WMU training in adults with neurological disorders, the small number of relevant studies conducted in that population to date precluded our ability to draw any meaningful conclusions. A brief examination of initial reports, however, suggests there is a potential benefit. We would like to emphasize the imperative for further neuroimaging studies with a pretest-posttest control group design involving adults with brain damage. There is an urgent need to develop and validate training interventions for neurological populations in order to establish an optimal training protocol and ultimately translate research findings into a clinical setting.

Acknowledgments

We acknowledge grant support by the Neurosciences Foundation, Margaret Murdoch Charitable Trust and Dr Mortimer and Theresa Sackler Foundation. The funding bodies had no role in the writing of the report or in the decision to submit the article for publication. We are grateful to Dr Simone Kühn and Dr Mónica Emch for providing summary measures of variability and data on training duration, respectively; essential for completing the meta-analyses.

Appendix A: Supplementary Material

Figures Captions

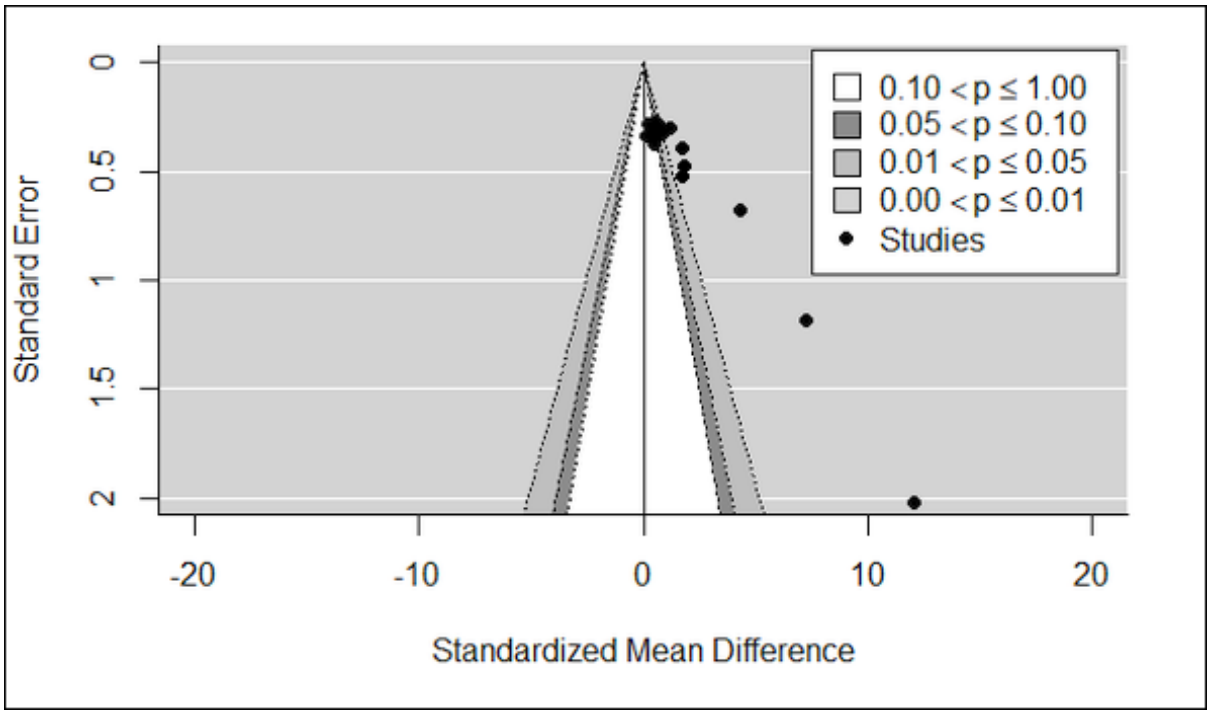


Figure S1: Contour-enhanced funnel plot for the overall training effect. The funnel is centered at 0 where the studies concentrating around the midline have no significant effects. The data points falling outside and to the bottom right of the funnel tend to have smaller sample sizes and large variance, in addition to significant and large effect sizes, and thus are more likely to bias the overall effect. The Egger's regression test for funnel plot asymmetry yielded significant results ($z = 9.36$, $p < .0001$) further corroborating the assumption for publication bias.

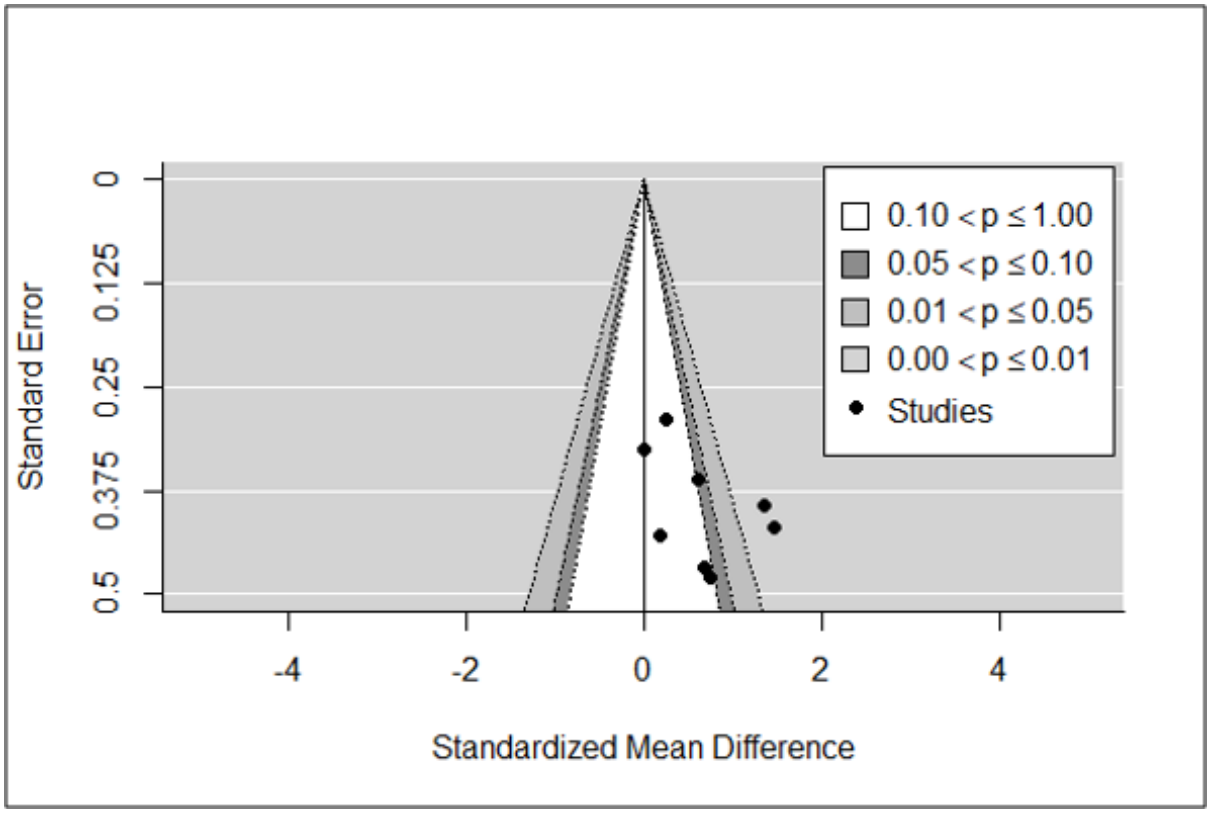


Figure S2: Contour-enhanced funnel plot for the near transfer effect. The dispersion of data points in the funnel indicate asymmetry, although the Egger's regression test proved non-significant in this case (z = 1.30, p = 0.19). Only two studies exhibit significant near transfer effects, suggesting that publication bias is unlikely the cause for such asymmetry.

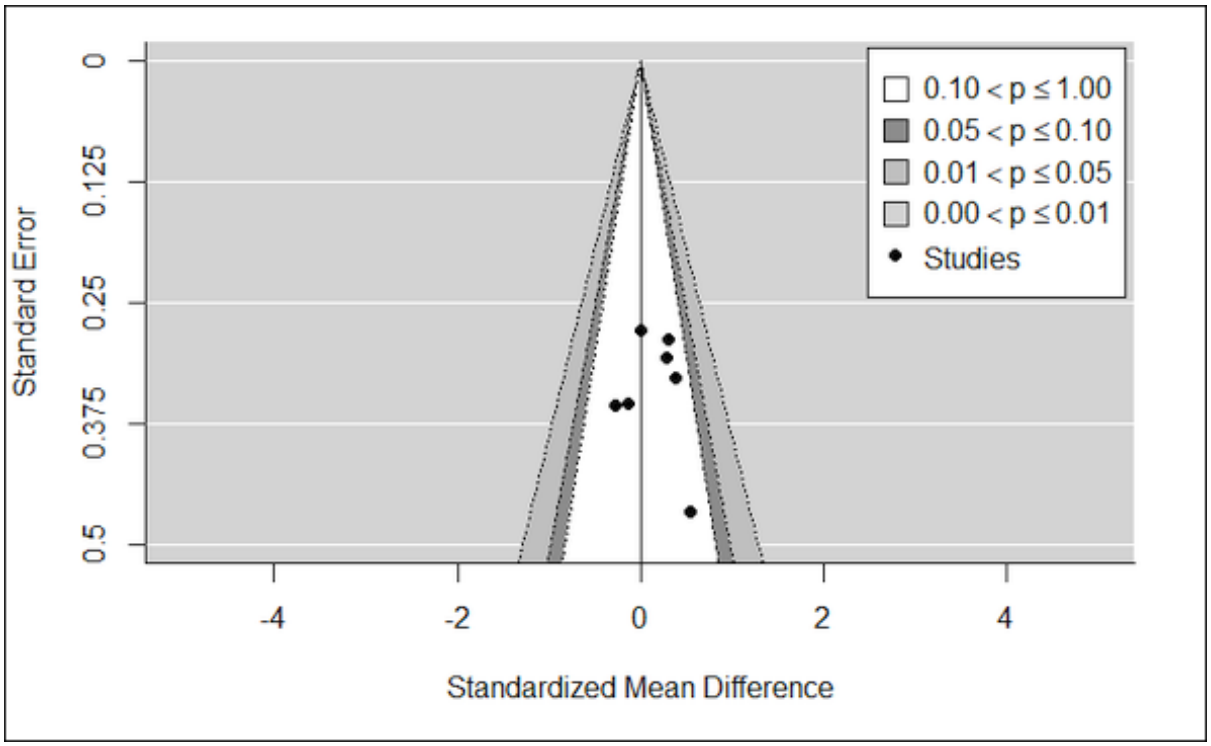


Figure S3: Contour-enhanced funnel plot for the far transfer effect. The data points do not indicate asymmetry, however the small number of studies testing for far transfer makes it

difficult to draw conclusions. None of the studies exhibited significant effect sizes and the Egger's regression test yielded non-significant results ($z = 0.26$, $p = 0.79$).

Tables

Table S1: PEDro-P Quality Assessment for reviewed studies

Reference	***Q1***	***Q2***	***Q3***	***Q4***	***Q5***	***Q6***	***Q7***	***Q8***	***Q9***	***Q10***	***Q11***	***Total***	***Quality Rating***
	Healthy Adults												
Aguirre et al. (51)	1	1	0	1	0	1	0	1	0	0	1	5	fair
Backman et al. (52)	1	1	0	1	0	0	0	0	0	1	1	4	fair
Backman et al. (53)	1	1	0	1	0	0	0	1	0	1	1	5	fair
Biel et al. (54)	1	1	0	1	1	1	0	1	0	1	0	6	good
Buschkuehl et al. (77)	1	0	0	1	0	1	0	0	0	1	1	4	fair
Clark et al. (73)	1	1	1	1	1	1	1	0	0	1	1	8	good
Colom et al.(56)	1	0	0	1	0	0	0	0	0	1	0	2	poor
Colom et al. (57)	1	0	0	1	0	0	0	0	0	1	0	2	poor
Dahlin et al. (37)	1	1	0	1	0	0	0	1	0	1	1	4	fair
Emch et al. (58)	1	0	0	1	1	0	0	1	0	1	1	5	fair
Finc et al. (59)	1	1	0	1	1	0	1	1	0	1	1	7	good
Flegal et al. (78)	1	1	1	1	1	1	0	0	0	1	1	7	good
Heinzel et al. (60)	1	0	0	0	0	0	0	1	1	0	1	3	poor
Heinzel et al. (61)	1	0	0	1	0	0	0	1	0	1	1	4	fair
Heinzel et al. (62)	1	0	0	0	0	0	0	1	1	0	1	3	poor
Hempel et al. (63)	1	0	0	0	0	0	0	0	0	0	1	1	poor
Kuhn et al. (72)	1	1	0	1	1	1	0	0	0	1	0	5	fair
Lawlor Savage et al. (74)	1	1	1	1	1	1	1	0	0	1	1	8	good
Miro-Padilla et al. (64,65) [†]	1	1	0	1	0	0	0	0	0	1	1	4	fair
Opitz et al. (66)	1	0	0	1	0	0	0	0	0	1	1	3	poor
Roman et al. (68)	1	0	0	1	0	0	0	0	0	1	1	3	poor
Roman et al.(67)	1	0	0	1	0	0	0	0	0	1	1	3	poor
Salminen et al.(69)	1	0	0	1	0	0	0	1	0	1	1	4	fair
Schneiders et al. (70)	1	0	0	1	0	0	0	0	0	1	1	3	poor

Schneiders et al. (80)	1	0	0	1	0	0	0	0	0	1	1	3	poor
Schweizer et al. (71)	1	1	0	0	0	1	0	1	0	1	1	5	fair
Thompson et al. (79)	1	1	1	1	0	0	0	0	0	1	1	5	fair
Neurological Populations													
Aguirre et al. (51)	1	1	0	1	0	1	0	1	0	0	1	5	fair
Bonzano et al. (55)	1	0	0	0	0	0	0	1	0	0	1	2	poor
Leung et al. (75)	0	0	0	0	0	0	0	1	1	0	0	2	poor
Leung et al. (76)	1	0	0	0	0	0	0	1	1	0	0	2	poor

Quality Rating: Good: score ≥ 6, Fair: score of 4-5 and Poor: score ≤ 3, **Healthy Adults: Mean:** 4.29, **Median:** 4, **SD:** 1.74, **Neurological Populations: Mean:** 2.75, **Median:** 2, **SD:** 1.50, **Q1** did not count towards the total score. **†**These studies share the same dataset; **Q1. Eligibility** criteria were specified, **Q2.** Subjects were randomly allocated to interventions (in a crossover study, subjects were randomly allocated an order in which treatments were received), **Q3.** Allocation was concealed, **Q4**. the intervention groups were similar at baseline regarding the most important prognostic indicators, **Q5**. There was blinding of all subjects, **Q6. There** was blinding of all therapists who administered the therapy, **Q7.** There was blinding of all assessors who measured at least one key outcome, **Q8.** Measures of at least one key outcome were obtained from more than 85% of the subjects initially allocated to groups, **Q9.** All subjects for whom outcome measures were available received the treatment or control condition as allocated or, where this was not the case, data for at least one key outcome was analysed by "intention to treat", **Q10.** The results of between-intervention group statistical comparisons are reported for at least one key outcome, **Q11.** The study provides both point measures and measures of variability for at least one key outcome.

Table S2*:* MRI Functional Connectivity Changes after WMU training in Healthy Adult Studies.

Reference	***Functional Connectivity Changes***	
	Training	***Transfer***
Finc et al. (59)	*Session by Group comparisons* *Whole-Brain Modularity increases, ns* ($\chi^2(1) = 1.50$, $p = 0.68$). TG showed a higher network modularity (M=3.09) compared to the CG(M=2.87). *Dynamic reorganization of large-scale systems* *Recruitment* - *Frontoparietal System:* ($\chi^2(3) = 9.03$, $p = 0.028$. ↑**Increase in recruitment** for the TG compared to the CG post training. The largest increase was reported for the TG between pre and post-training t(120)= −2.892, p=0.027, Bonferroni-corrected). No significant changes for the CG, t(120)=−1.169, p=1. - *Default mode system,* ns, ($\chi^2(3) = 2.66$, $p = 0.48$). ↑**Increased** recruitment higher for TG compared to CG, t(165.6)=−3.03, p=0.003). *Integration of systems* - *Frontoparietal with default mode systems*, ($\chi^2(3) = 14.25$, p=0.0025). ↓***Decrease*** post-training *only found for the TG compared to the CG (*t(120)= 4.37, p=0.0002,*).* - *Subcortical* with dorsal attention, ventral attention, cingulo-opercular and auditory systems: ↑**Increase** at early training stages and a ↓**decrease** later on. - *Subcortical with default-mode systems:* initial ↓**decrease** at early training stages and an ↑**increase** at later stages	-
Heinzel et al. (60)	*session by Load comparison*	-

	No differences in connectivity changes in the WM network.	
Thompson et al. (79)	*Session by Group comparison* ↑**Increased functional connectivity for the TG** was observed for all 4 pairings of prefrontal and parietal ROIs in the 2-back condition ($p < 0.05$, Bonferroni). No changes for 1- or 3-back.	-

Table S3: MRI Structural Changes after WMU training in Healthy Adult Studies.

Reference	***Structural Changes***	***Structural Connectivity Changes***
Biel et al. (54)	*Group by time comparisons, ns* (FWE, p<0.05 whole-brain). No changes in GM volume, myelination and iron levels.	-
*Colom et al. (56)***	Voxel-based independent samples t-tests, TG Vs CG post-training. Significant change post-training in L/R temporal lobe. ↓**Decreased volume** in the CG **- Volume preservation** in the TG.	-
*Colom et al., (57) ****	*Group by Time comparisons* ↑ ***Increased regional gray matter volume*** *for the TG post-training in:* i. L posterior cingulate cortex ii. R cerebellum iii. R temporal lobe	-
Heinzel et al. (60)	No significant changes in the WM network GM volume after training (t(14)=0.83, p=0.421).	-
Lawlor-Savage et al. (74)	*Group By time comparisons, alpha <.001* **No significant effects** for cortical surface area, thickness or volume changes in any of the frontal, parietal lobe regions of interest, in cingulate or insular cortical regions or volume estimates within subcortical regions of interest. **No significant effects** for total subcortical GM volumes or total GM volumes.	-
*Roman et al. (68)****	Mean cortical thickness (CT) and cortical surface area (CSA) were computed at each ROI for the TG and CG before and after training. The standardised change was computed. **ANCOVA, CT differences between TG and CG in:** i. R Ventral frontal cortex ii. R Middle temporal cortex	-

	Minor thickening for TG. **Minor thinning** for CG. **ANCOVA, CSA differences between TG and CG in:** i. R pas opercularis ii. R posterolateral temporal cortex **Expanding effect** for the TG. **Contracting effect** for the CG.	
*Roman et al. (67)****	Changes more pronounced in the TG compared to the CG in the Sub-Network containing temporal, frontal, parietal, subcortical regions and the insula. **Connectome Topological Properties** Time by Group comparison for the sub-network ↑ **Increase** in **Global efficiency** (**Eg)** for the TG. ↑ **Increase** in **Strength (S)** for the TG. **No change in CG for either Eg or S.**	Network-Based Statistics to identify connectional sub-networks modulated by cognitive training. Changes more pronounced in the TG compared to the CG in the Sub-Network containing temporal, frontal, parietal, subcortical regions and the insula. Most highly connected node in this network was located in the **L middle temporal region** and was highly interconnected with: i. L/R basal forebrain ii. L parahippocampal area iii. L pallidum iv. L supramraginal v. L inferior parietal area vi. R insula vii. R accumbens viii. R post central gyrus ix. R pars opercularis ↑ **Increase in Connectivity** for the TG in this network. No changes for CG.

*** These studies shared the same Bx dataset, **Bx:** behavioural, **FWE:** Family-Wise Error, **GM:** Gray Matter.

References

1. Kaas JH. Neural Plasticity. In: Smelser NJ, Baltes PB, editors. International Encyclopedia of the Social & Behavioral Sciences. Oxford: Pergamon; 2001. p. 10542-6. https://doi.org/10.1016/B0-08-043076-7/03619-6

2. Lovden M, Backman L, Lindenberger U, Schaefer S, Schmiedek F. A theoretical framework for the study of adult cognitive plasticity. Psychol Bull. 2010;136(4):659-76. https://doi.org/10.1037/a0020080

3. Baddeley A. Working memory. Science. 1992;255(5044):556. https://doi.org/10.1126/science.1736359

4. Miyake A, Friedman NP, Emerson MJ, Witzki AH, Howerter A, Wager TD. The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: a latent variable analysis. Cogn Psychol. 2000;41(1):49-100. https://doi.org/10.1006/cogp.1999.0734

5. Nee DE, Brown JW, Askren MK, Berman MG, Demiralp E, Krawitz A, et al. A meta-analysis of executive components of working memory. Cereb Cortex. 2013;23(2):264-82. https://doi.org/10.1093/cercor/bhs007

6. Collette F, Hogge M, Salmon E, Van der Linden M. Exploration of the neural substrates of executive functioning by functional neuroimaging. Neuroscience. 2006;139(1):209-21. https://doi.org/10.1016/j.neuroscience.2005.05.035

7. Wager TD, Smith EE. Neuroimaging studies of working memory. Cognitive, Affective, & Behavioral Neuroscience. 2003;3(4):255-74. https://doi.org/10.3758/CABN.3.4.255

8. Burgess PW, Stuss DT. Fifty Years of Prefrontal Cortex Research: Impact on Assessment. Journal of the International Neuropsychological Society. 2017;23(9-10):755-67. https://doi.org/10.1017/S1355617717000704

9. Diamond A. Executive Functions. Annual Review of Psychology. 2013;64(1):135-68. https://doi.org/10.1146/annurev-psych-113011-143750

10. Shallice TIM, Burgess PW. Deficits in strategy application following frontal lobe damage in man. Brain. 1991;114(2):727-41. https://doi.org/10.1093/brain/114.2.727

11. Emch M, von Bastian CC, Koch K. Neural Correlates of Verbal Working Memory: An fMRI Meta-Analysis. Frontiers in Human Neuroscience. 2019;13(180). https://doi.org/10.3389/fnhum.2019.00180

12. Salmi J, Nyberg L, Laine M. Working memory training mostly engages general-purpose large-scale networks for learning. Neuroscience and Biobehavioral Reviews. 2018; 93:108-22. https://doi.org/10.1016/j.neubiorev.2018.03.019

13. Hsu NS, Novick JM, Jaeggi SM. The development and malleability of executive control abilities. Frontiers in Behavioral Neuroscience. 2014;8(JUNE):221. https://doi.org/10.3389/fnbeh.2014.00221

14. Klingberg T. Training and plasticity of working memory. Trends Cogn Sci. 2010;14(7):317-24. https://doi.org/10.1016/j.tics.2010.05.002

15. Dougherty MR, Hamovitz T, Tidwell JW. Reevaluating the effectiveness of n-back training on transfer through the Bayesian lens: Support for the null. Psychon B Rev. 2016;23(1):306-16. https://doi.org/10.3758/s13423-015-0865-9

16. Melby-Lervåg M, Hulme C. Is working memory training effective? A meta-analytic review. Developmental Psychology. 2013;49(2):270-91.https://doi.org/10.1037/a0028228

17. Melby-Lervåg M, Redick TS, Hulme C. Working Memory Training Does Not Improve Performance on Measures of Intelligence or Other Measures of "Far Transfer": Evidence From a Meta-Analytic Review. Perspectives on Psychological Science. 2016;11(4):512-34. https://doi.org/10.1177/1745691616635612

18. Soveri A, Antfolk J, Karlsson L, Salo B, Laine M. Working memory training revisited: A multi-level meta-analysis of n-back training studies. Psychon B Rev. 2017;24(4):1077-96. https://doi.org/10.3758/s13423-016-1217-0

19. Brehmer Y, Rieckmann A, Bellander M, Westerberg H, Fischer H, Backman L. Neural correlates of training-related working-memory gains in old age. NeuroImage. 2011;58(4):1110-20. https://doi.org/10.1016/j.neuroimage.2011.06.079

20. Dahlin E, Backman L, Neely AS, Nyberg L. Training of the executive component of working memory: subcortical areas mediate transfer effects. Restor Neurol Neurosci. 2009;27(5):405-19. https://doi.org/10.3233/RNN-2009-0492

21. Au J, Sheehan E, Tsai N, Duncan GJ, Buschkuehl M, Jaeggi SM. Improving fluid intelligence with training on working memory: a meta-analysis. Psychon Bull Rev. 2015;22(2):366-77. https://doi.org/10.3758/s13423-014-0699-x

22. Au J, Gibson BC, Bunarjo K, Buschkuehl M, Jaeggi SM. Quantifying the Difference Between Active and Passive Control Groups in Cognitive Interventions Using Two Meta-analytical Approaches. Journal of Cognitive Enhancement. 2020;4(2):192-210. https://doi.org/10.1007/s41465-020-00164-6

23. von Bastian CC, Oberauer K. Effects and mechanisms of working memory training: a review. Psychol Res. 2014;78(6):803-20. https://doi.org/10.1007/s00426-013-0524-6

24. Kelly C, Foxe JJ, Garavan H. Patterns of Normal Human Brain Plasticity After Practice and Their Implications for Neurorehabilitation. Archives of Physical Medicine and Rehabilitation. 2006;87(12 SUPPL.):20-9. https://doi.org/10.1016/j.apmr.2006.08.333

25. Jolles DD, van Buchem MA, Crone EA, Rombouts SARB. Functional brain connectivity at rest changes after working memory training. Human Brain Mapping. 2013;34(2):396-406. https://doi.org/10.1002/hbm.21444

26. Takeuchi H, Taki Y, Nouchi R, Hashizume H, Sekiguchi A, Kotozaki Y, et al. Effects of working memory training on functional connectivity and cerebral blood flow during rest. Cortex. 2013;49(8):2106-25. https://doi.org/10.1016/j.cortex.2012.09.007

27. Lövdén M, Wenger E, Mårtensson J, Lindenberger U, Bäckman L. Structural brain plasticity in adult learning and development. Neuroscience & Biobehavioral Reviews. 2013;37(9, Part B):2296-310.https://doi.org/10.1016/j.neubiorev.2013.02.014

28. Zatorre RJ, Fields RD, Johansen-Berg H. Plasticity in gray and white: neuroimaging changes in brain structure during learning. Nature Neuroscience. 2012;15(4):528-36. https://doi.org/10.1038/nn.3045

29. Takeuchi H, Taki Y, Sassa Y, Hashizume H, Sekiguchi A, Fukushima A, et al. Working Memory Training Using Mental Calculation Impacts Regional Gray Matter of the Frontal and Parietal Regions. PLOS ONE. 2011;6(8):e23175. https://doi.org/10.1371/journal.pone.0023175

30. Metzler-Baddeley C, Caeyenberghs K, Foley S, Jones DK. Task complexity and location specific changes of cortical thickness in executive and salience networks after working memory training. NeuroImage. 2016;130:48-62. https://doi.org/10.1016/j.neuroimage.2016.01.007

31. Caeyenberghs K, Metzler-Baddeley C, Foley S, Jones DK. Dynamics of the Human Structural Connectome Underlying Working Memory Training. Journal of Neuroscience. 2016;36(14):4056-66. https://doi.org/10.1523/JNeurosci.1973-15.2016

32. Takeuchi H, Sekiguchi A, Taki Y, Yokoyama S, Yomogida Y, Komuro N, et al. Training of Working Memory Impacts Structural Connectivity. Journal of Neuroscience. 2010;30(9):3297-303. https://doi.org/10.1523/JNEUROSCI.4611-09.2010

33. Schmiedek F, Lovden M, Lindenberger U. A task is a task is a task: putting complex span, n-back, and other working memory indicators in psychometric context. Front Psychol. 2014;5:1475.https://doi.org/10.3389/fpsyg.2014.01475

34. Jaeggi SM, Buschkuehl M, Perrig WJ, Meier B. The concurrent validity of the N-back task as a working memory measure. Memory. 2010;18(4):394-412. https://doi.org/10.1080/09658211003702171

35. Linares R, Borella E, Lechuga MT, Carretti B, Pelegrina S. Nearest transfer effects of working memory training: A comparison of two programs focused on working memory updating. PLOS ONE. 2019;14(2):e0211321. https://doi.org/10.1371/journal.pone.0211321

36. Moher D, Liberati A, Tetzlaff J, Altman DG, The PG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLOS Medicine. 2009;6(7):e1000097. https://doi.org/10.1371/journal.pmed.1000097

37. Dahlin E, Neely AS, Larsson A, Backman L, Nyberg L. Transfer of learning after updating training mediated by the striatum. Science. 2008;320(5882):1510-2. https://doi.org/10.1126/science.1155466

38. Maher CG, Sherrington C, Herbert RD, Moseley AM, Elkins M. Reliability of the PEDro scale for rating quality of randomized controlled trials. Phys Ther. 2003;83(8):713-21. https://doi.org/10.1093/ptj/83.8.713

39. Van Criekinge T, Truijen S, Schröder J, Maebe Z, Blanckaert K, van der Waal C, et al. The effectiveness of trunk training on trunk control, sitting and standing balance and mobility post-stroke: a systematic review and meta-analysis. Clinical Rehabilitation. 2019;33(6):992-1002. https://doi.org/10.1177/0269215519830159

40. Review Manager (RevMan). Version 5.3. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2014.

41. Higgins JPT, Deeks JJ. Selecting Studies and Collecting Data. Cochrane Handbook for Systematic Reviews of Interventions2008. p. 151-85. https://doi.org/10.1002/9780470712184.ch7

42. Special Topics in Statistics. Cochrane Handbook for Systematic Reviews of Interventions. p. 481-529.

43. Zhang H, Huntley J, Bhome R, Holmes B, Cahill J, Gould RL, et al. Effect of computerised cognitive training on cognitive outcomes in mild cognitive impairment: a systematic review and meta-analysis. BMJ Open. 2019;9(8):e027062. https://doi.org/10.1136/bmjopen-2018-027062

44. Hedges LV, Olkin I. Statistical methods for meta-analysis. London;Orlando;: Academic Press; 1985.

45. Cohen J. Statistical power analysis for the behavioral sciences: Routledge; 2013. https://doi.org/10.4324/9780203771587

46. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. Bmj. 2003;327(7414):557-60. https://doi.org/10.1136/bmj.327.7414.557

47. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. 2010. 2010;36(3):48. https://doi.org/10.18637/jss.v036.i03

48. Team R. RStudio: Integrated Development Environment for R. RStudio, Inc., Boston, MA, 2019. Retrieved from http://www.rstudio.com/:

49. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2020. Retrieved from https://www.R-project.org/.

50. Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629-34. https://doi.org/10.1136/bmj.315.7109.629

51. Aguirre N, Cruz-Gomez AJ, Miro-Padilla A, Bueicheku E, Torres RB, Avila C, et al. Repeated Working Memory Training Improves Task Performance and Neural Efficiency in Multiple Sclerosis Patients and Healthy Controls. Multiple Sclerosis International. 2019;2019:2657902. https://doi.org/10.1155/2019/2657902

52. Bäckman L, Nyberg L, Soveri A, Johansson J, Andersson M, Dahlin E, et al. Effects of Working-Memory Training on Striatal Dopamine Release. Science. 2011;333(6043):718. https://doi.org/10.1126/science.1204978

53. Backman L, Waris O, Johansson J, Andersson M, Rinne JO, Alakurtti K, et al. Increased dopamine release after working-memory updating training: Neurochemical correlates of transfer. Scientific reports. 2017;7(1):7160. https://doi.org/10.1038/s41598-017-07577-y

54. Biel D, Steiger TK, Volkmann T, Jochems N, Bunzeck N. The gains of a 4-week cognitive training are not modulated by novelty. Human Brain Mapping. 2020. https://doi.org/10.1002/hbm.24965

55. Bonzano L, Pedulla L, Pardini M, Tacchino A, Zaratin P, Battaglia MA, et al. Brain activity pattern changes after adaptive working memory training in multiple sclerosis. Brain Imaging and Behavior. 2020;14(1):142-54. https://doi.org/10.1007/s11682-018-9984-z

56. Colom R, Hua X, Martinez K, Burgaleta M, Roman FJ, Gunter JL, et al. Brain structural changes following adaptive cognitive training assessed by Tensor-Based Morphometry (TBM). Neuropsychologia. 2016;91:77-85. https://doi.org/10.1016/j.neuropsychologia.2016.07.034

57. Colom R, Martínez K, Burgaleta M, Román FJ, García-García D, Gunter JL, et al. Gray matter volumetric changes with a challenging adaptive cognitive training program based on the dual n-back task. Personality and Individual Differences. 2016;98:127-32. https://doi.org/10.1016/j.paid.2016.03.087

58. Emch M, Ripp I, Wu Q, Yakushev I, Koch K. Neural and Behavioral Effects of an Adaptive Online Verbal Working Memory Training in Healthy Middle-Aged Adults. Frontiers in aging neuroscience. 2019;11:300. https://doi.org/10.3389/fnagi.2019.00300

59. Finc K, Bonna K, He X, Lydon-Staley DM, Kuhn S, Duch W, et al. Dynamic reconfiguration of functional brain networks during working memory training. Nature communications. 2020;11(1):2435. https://doi.org/10.1038/s41467-020-15631-z

60. Heinzel S, Lorenz RC, Brockhaus WR, Wüstenberg T, Kathmann N, Heinz A, et al. Working memory load-dependent brain response predicts behavioral training gains in older adults. Journal of Neuroscience. 2014;34(4):1224-33. https://doi.org/10.1523/JNEUROSCI.2463-13.2014

61. Heinzel S, Lorenz RC, Pelz P, Heinz A, Walter H, Kathmann N, et al. Neural correlates of training and transfer effects in working memory in older adults. NeuroImage. 2016;134:236-49. https://doi.org/10.1016/j.neuroimage.2016.03.068

62. Heinzel S, Rimpel J, Stelzel C, Rapp MA. Transfer Effects to a Multimodal Dual-Task after Working Memory Training and Associated Neural Correlates in Older Adults - A Pilot Study. Frontiers in human neuroscience. 2017;11:85. https://doi.org/10.3389/fnhum.2017.00085

63. Hempel A, Giesel FL, Garcia Caraballo NM, Amann M, Meyer H, Wüstenberg T, et al. Plasticity of Cortical Activation Related to Working Memory during Training. American Journal of Psychiatry. 2004;161(4):745-7.https://doi.org/10.1176/appi.ajp.161.4.745

64. Miro-Padilla A, Bueicheku E, Avila C. Locating neural transfer effects of n-back training on the central executive: a longitudinal fMRI study. Scientific reports. 2020;10(1):5226. https://doi.org/10.1038/s41598-020-62067-y

65. Miro-Padilla A, Bueicheku E, Ventura-Campos N, Flores-Compan MJ, Parcet MA, Avila C. Long-term brain effects of N-back training: an fMRI study. Brain Imaging and Behavior. 2018;13(4):1-13. https://doi.org/10.1007/s11682-018-9925-x

66. Opitz B, Schneiders JA, Krick CM, Mecklinger A. Selective transfer of visual working memory training on Chinese character learning. Neuropsychologia. 2014;53(1):1-11. https://doi.org/10.1016/j.neuropsychologia.2013.10.017

67. Roman FJ, Iturria-Medina Y, Martinez K, Karama S, Burgaleta M, Evans AC, et al. Enhanced structural connectivity within a brain sub-network supporting working memory and engagement processes after cognitive training. Neurobiology of Learning and Memory. 2017;141:33-43. https://doi.org/10.1016/j.nlm.2017.03.010

68. Roman FJ, Lewis LB, Chen C-H, Karama S, Burgaleta M, Martinez K, et al. Gray matter responsiveness to adaptive working memory training: a surface-based morphometry study. Brain structure & function. 2016;221(9):4369-82. https://doi.org/10.1007/s00429-015-1168-7

69. Salminen T, Kuhn S, Frensch PA, Schubert T. Transfer after dual n-back training depends on striatal activation change. Journal of Neuroscience. 2016;36(39):10198-213. https://doi.org/10.1523/JNEUROSCI.2305-15.2016

70. Schneiders JA, Opitz B, Krick CM, Mecklinger A. Separating intra-modal and across-modal training effects in visual working memory: An fMRI investigation. Cerebral Cortex. 2011;21(11):2555-64. https://doi.org/10.1093/cercor/bhr037

71. Schweizer S, Grahn J, Hampshire A, Mobbs D, Dalgleish T. Training the emotional brain: improving affective control through emotional working memory training. Journal of neuroscience. 2013;33(12):5301-11. https://doi.org/10.1523/JNEUROSCI.2593-12.2013

72. Kuhn S, Schmiedek F, Noack H, Wenger E, Bodammer NC, Lindenberger U, et al. The dynamics of change in striatal activity following updating training. Human Brain Mapping. 2013;34(7):1530-41. https://doi.org/10.1002/hbm.22007

73. Clark CM, Lawlor-Savage L, Goghari VM. Functional brain activation associated with working memory training and transfer. Behavioural Brain Research. 2017;334:34-49. https://doi.org/10.1016/j.bbr.2017.07.030

74. Lawlor-Savage L, Clark CM, Goghari VM. No evidence that working memory training alters gray matter structure: A MRI surface -based analysis. Behavioural Brain Research. 2019;360:323-40. https://doi.org/10.1016/j.bbr.2018.12.008

75. Leung A, Ng B, Yuen J, Dixon G, Kim E. Neuroplastic Changes after Auditory Working Memory Training in a Patient Surviving Multiple Strokes. Journal of Neuroscience and Rehabilitation. 2014;1(1):1-22. https://doi.org/10.17653/2374-9091.SS0002

76. Leung AWS, Barrett LM, Butterworth D, Werther K, Dawson DR, Brintnell ES. Neural Plastic Effects of Working Memory Training Influenced by Self-perceived Stress in Stroke: A Case Illustration. Frontiers in psychology. 2016;7:1266. https://doi.org/10.3389/fpsyg.2016.01266

77. Buschkuehl M, Hernandez-Garcia L, Jaeggi SM, Bernard JA, Jonides J. Neural effects of short-term training on working memory. Cognitive, affective & behavioral neuroscience. 2014;14(1):147-60. https://doi.org/10.3758/s13415-013-0244-9

78. Flegal KE, Ragland JD, Ranganath C. Adaptive task difficulty influences neural plasticity and transfer of training. NeuroImage. 2019;188:111-21. https://doi.org/10.1016/j.neuroimage.2018.12.003

79. Thompson TW, Waskom ML, Gabrieli JDE. Intensive working memory training produces functional changes in large-scale frontoparietal networks. Journal of cognitive neuroscience. 2016;28(4):575-88. https://doi.org/10.1162/jocn_a_00916

80. Schneiders JA, Opitz B, Tang H, Deng Y, Xie C, Li H, et al. The impact of auditory working memory training on the fronto-parietal working memory network. Frontiers in Human Neuroscience. 2012;(JUNE 2012) (no pagination)(173):173. https://doi.org/10.3389/fnhum.2012.00173

81. Schott BH, Minuzzi L, Krebs RM, Elmenhorst D, Lang M, Winz OH, et al. Mesolimbic Functional Magnetic Resonance Imaging Activations during Reward Anticipation Correlate with

Reward-Related Ventral Striatal Dopamine Release. The Journal of Neuroscience. 2008;28(52):14311. https://doi.org/10.1523/JNEUROSCI.2058-08.2008

82. Cabeza R. Hemispheric Asymmetry Reduction in Older Adults: The HAROLD Model. Psychology and Aging. 2002;17(1):85-100. https://doi.org/10.1037/0882-7974.17.1.85

83. Grady CL, Maisog JM, Horwitz B, Ungerleider LG, Mentis MJ, Salerno JA, et al. Age-related changes in cortical blood flow activation during visual processing of faces and location. The Journal of Neuroscience. 1994;14(3):1450. https://doi.org/10.1523/JNEUROSCI.14-03-01450.1994

84. Reuter-Lorenz PA, Jonides J, Smith EE, Hartley A, Miller A, Marshuetz C, et al. Age Differences in the Frontal Lateralization of Verbal and Spatial Working Memory Revealed by PET. Journal of Cognitive Neuroscience. 2000;12(1):174-87. https://doi.org/10.1162/089892900561814

85. Reuter-Lorenz PA, Cappell KA. Neurocognitive Aging and the Compensation Hypothesis. Current Directions in Psychological Science. 2008;17(3):177-82. https://doi.org/10.1111/j.1467-8721.2008.00570.x

86. Iordan AD, Cooke KA, Moored KD, Katz B, Buschkuehl M, Jaeggi SM, et al. Neural correlates of working memory training: Evidence for plasticity in older adults. NeuroImage. 2020;217:116887. https://doi.org/10.1016/j.neuroimage.2020.116887

87. Doyon J, Benali H. Reorganization and plasticity in the adult brain during learning of motor skills. Curr Opin Neurobiol. 2005;15(2):161-7. https://doi.org/10.1016/j.conb.2005.03.004

88. Lustig C, Shah P, Seidler R, Reuter-Lorenz PA. Aging, training, and the brain: a review and future directions. Neuropsychol Rev. 2009;19(4):504-22. https://doi.org/10.1007/s11065-009-9119-9

89. Nyberg L, Eriksson J, Medicinska f, Umeå centrum för funktionell h, Umeå u. Working Memory: Maintenance, Updating, and the Realization of Intentions. COLD SPRING HARBOR PERSPECTIVES IN BIOLOGY. 2016;8(2):a021816. https://doi.org/10.1101/cshperspect.a021816

90. Cools R, D'Esposito M. Inverted-U-Shaped Dopamine Actions on Human Working Memory and Cognitive Control. Biological Psychiatry. 2011;69(12):e113-e25. https://doi.org/10.1016/j.biopsych.2011.03.028

91. Colom R, Román FJ, Abad FJ, Shih PC, Privado J, Froufe M, et al. Adaptive n-back training does not improve fluid intelligence at the construct level: Gains on individual tests suggest that training may enhance visuospatial processing. Intelligence. 2013;41(5):712-27. https://doi.org/10.1016/j.intell.2013.09.002

92. Thompson TW, Waskom ML, Garel KL, Cardenas-Iniguez C, Reynolds GO, Winter R, et al. Failure of working memory training to enhance cognition or intelligence. PLoS One. 2013;8(5):e63614.https://doi.org/10.1371/journal.pone.0063614

Figure Captions

Figure 1: Summary of Study Identification & Selection.

Figure 2: Training effect meta-analysis: Active & Passive CG sub-group analyses. One study (69) involved both an ACG and a PCG, hence they were included in both sub-group analyses. [1] Experiment 1: young adults, [2] Experiment 2: older adults.

Figure 3: Training effect meta-analysis: Shorter duration and longer duration sub-group analyses. In this analysis, the ACG and PCG for the study that involved both (69) were combined into one CG, hence its training effect size is different to that reported in Figure 2. For the same reason, the total N value for the TG differs between Figures 2 and 3. Consequently there is a very small difference in the total overall effect between these analyses. [1] Experiment 1: young adults, [2] Experiment 2: older adults.

*Figure 4***:** Relationship between training hours and training effect for Active & Passive CG sub-group comparisons.

Figure 5: Transfer effect meta-analysis: A. Near transfer after WMU training, B. Far transfer after WMU training. [1] Experiment 1: young adults. [2] Experiment 2: older adults.

Figure 6: Schematic model for dynamic activity changes determined by repeated exposure to training and transfer tasks.

Tables

Table 1: Study and training characteristics for reviewed studies

Reference	***Study Sample (N, Age mean ± SD years)***	***Study Design***	***Total Training Hours, no of sessions pw (Weeks total, Sessions total, Minutes per session)***	***Training (modality) (difficulty)***	***Control Group (control task)***
Healthy Adults					
Aguirre et al. (51)	Healthy Adults (N=29, 32.72±7.48)* - TG (N=14, 31.21 ±8.72), - PCG(N=15, 34.13±6.07)	Randomised controlled trial	4 hours, 4 sessions pw (1 week, 4 sessions, 60 min per session)	Single N-back *(verbal, adaptive)*	Passive
Backman et al. (52)	Healthy Adults (N=20, 22.25±3.17*) - TG (N=10, 22.8 ±3.9), - PCG (N=10, 21.7 ±2.3	Randomised controlled trial	11.24 hours, 3 sessions pw (5 weeks,15 sessions, 45 min per session)	1. Letter Memory updating 2. Number updating 3. Letter updating 4. Colour updating 5. Spatial location updating 6. Verbal Keep Track *(all adaptive)*	Passive
Backman et al. (53)	Healthy Adults (N = 27, 22.49 ± 1.61*) - TG (N=14, 22.21, ±1.72), - PCG (N=13, 22.79 ±1.48)	Randomised controlled trial	11.24 hours, 3 sessions pw (5 weeks, 15 sessions, 45 min per session)	1. Letter Memory updating 2. Number updating 3. Letter updating 4. Colour updating 5. Spatial location updating 6. Verbal Keep Track *(all adaptive)*	Passive

Biel et al. (54)	Healthy Adults (N = 83, 63.93±8.54*)
	- TG (N=56, 64.24± 8.85)*,
	  - TG1 (N = 28, 64.29 ± 9.69)
	  - TG2 (N=28, 64.18 ± 8.10)
	- PCG (N= 27, 63.30± 7.99)
		Randomised controlled trial	7.2 hours, 3 sessions pw (4 weeks, 12 sessions, 36 min per session)	TG1: Single 2-back + Novel nature movies (NOV) *(numerical, non-adaptive)*
				
				TG2: Single 2-back + Familiarised Nature movies (FAM), *(numerical, non-adaptive)*
					Passive

Buschkuehl et al. (77)	Healthy adults (N= 55, 21.8 ±2.7) - TG (N=27, 22.3± 3.1), - ACG (N= 28 , 21.2 ±2.1)	Quasi-experimental	2.5 hours, 7 sessions a week (1 week,7 sessions, 20 min per session)	Single N-back *(visuo-spatial, fixed*)	Active (Vocabulary & General Knowledge Questions)
*Clark et al. (73)***	Healthy Adults (N=76, 31.11±5.80) * - TG (N=25, 30.68 ± 6.24), - ACG (N=24, 31.33 ±5.78), - PCG(N=27, 31.32 ±5.58) N=49 were scanned from TG & ACG	Randomised controlled trial	10 hours, 5 sessions pw (6 weeks, 30 sessions, 20 min per session)	Lumosity Training 1. Memory Match (Single 2-back, *visual, fixed)* 2. Memory Match overload (Single 3-back, *visual, fixed)* 3. Memory Lane (Dual N-back, *adaptive)*	I. Active Lumosity Training 1.Processing Speed Speed Match (speeded Single 1-back Task, *visual*) 2.Speed Match overdrive (like Speed Match including partial match option, *visual)* 3.Spatial Speed Match (like the Speed match task but stimuli differ in spatial orientation) II. Passive
*Colom et al. (56) *** (Bx Data taken from*	Healthy Adults (N=56, 18.3±1.1) - TG (N=28, 18.04±0.9) - PCG (N=28, 18.2±1.2)	Quasi-experimental	12 hours, 2 sessions pw (12 weeks,24 sessions, 30 min per session)	Dual N-back *(auditory and visual, adaptive)*	Passive

Colom et al. (91))					
*Colom et al. (57)***, Bx (Data taken from Colom et al.(91)*	Healthy Adults (N=56, 18.12±1.05) * - TG (N=28, 18.04±0.9), - PCG (N=28, 18.2±1.2)	Quasi-experimental	12 hours, 2 sessions pw (12 weeks, 24 sessions, 30 min per session)	Dual N-back *(auditory and visual, adaptive)*	Passive
Dahlin et al. (37)	Healthy Young Adults (N=22, 23.59±2.48) * - TG (N= 15, 23.67±2.92), - PCG (N=7, 23.43±1.27) Healthy Older Adults (N=19, 68.32±1.79) * - TG (N=11, 68.27±1.79), - PCG (N=8, 68.38 ± 1.92)	Randomised controlled trial	11.25 hours, 3 sessions pw (5 weeks,15 sessions, 45 min per session)	1. Letter Memory updating 2. Number updating 3. Letter updating 4. Colour updating 5. Spatial location updating 6. Verbal Keep Track *(all adaptive)*	Passive
Emch et al. (58)	Healthy Older Adults (N=57, 55.85±4.24) - TG (N=30, 5.80±4.30), - ACG (N=27, 55.92 ± 4.25)	Quasi-experimental	10.66 hours, 4 sessions pw (8 weeks, 32 sessions, 20 min per session)	Single N-back *(verbal, adaptive)*	Active NA Single 1-back
Finc et al. (59)	Healthy Adults (N=53, 21.17, age range 18 to 28 years, SD = 2.5*) N = 46 were scanned - TG (N = 23) - ACG (N = 23)	Randomised controlled trial (matched by sex)	9 hours, 3 sessions pw (6 weeks, 18 sessions, 30 min per session)	Dual N-back *(auditory and visual, adaptive)*	Active Single N-back *(visual and auditory)*

Flegal et al. (78)	Healthy young adults (N=56, 20.8 ± 2.4) - TG (N=19, 20.32±1.73) - ACG (N=19, 20.79±2.92) - PCG (N=18, 21.33±2.20) N=38 were scanned from TG & ACG	Randomised controlled trial	8.33 hours, 4 sessions pw (3 weeks, 10 sessions, 50 min per session)	1. Matrix updating (*visuospatial)* 2. Verbal Keep Track *(all adaptive)*	I. Active 1.NA Matrix updating (*visuospatial)* 2.NA Verbal Keep Track II. Passive
*Heinzel et al. (60)*****	Healthy Older Adults (N= 19, 65.95±3.73) N=15 were scanned	Quasi experimental Single group	9 hours, 3 sessions pw (4 weeks, 12 sessions, 45 min per session)	Single N-back (*numerical, adaptive)*	No CG
Heinzel et al.(61) ****	Healthy Older Adults (N=29, 66.02±4.35) - TG (N=15, 66.04±4.04), - PCG (N=14, 66.00±4.82)	Quasi-experimental	9 hours, 3 sessions pw (4 weeks, 12 sessions, 45 min per session)	Single N-back (*numerical*, adaptive*)*	Passive
Heinzel et al. (62) ****	Healthy Older adults (N=38), final sample N=34 (range 60-70 years) - TG (N=18, 65,78±3.04) - PCG (N=16, 65 ±3.67) N=15 were scanned	Quasi-experimental Single group	9 hours, 3 sessions pw (4 weeks, 12 sessions, 45 min per session)	Single N-back (*numerical, adaptive)*	No CG
Hempel et al. (63)	Healthy Adults (N=9, age range 26 to 32, SD= 1.5) *	Quasi-experimental Single group	No information	Single N-back (*visuospatial, no information)*	No CG
Kuhn et al. (72)	Healthy Adults (N=46, 25.0±2.7) - TG (N=26, 24.7±2.3) - ACG (N=20 (25.4 ±3.1)	Quasi-experimental	27.65 hours (no info, 55 sessions, 31.5 min per session)	1. Number Memory Updating 2. Single N-back *(spatial) (all adaptive)*	Active 1. NA Number Memory Updating 2. NA N-back *(spatial)*

*Lawlor-Savage et al. (74)***	Healthy Adults (N=76, 31.11±5.80)* - TG (N=25, 30.68 ± 6.24), - ACG (N=24, 31.33 ±5.78), - PCG(N=27, 31.32 ±5.58) N=49 were scanned	Randomised controlled trial	10 hours, 5 sessions pw (6 weeks, 30 sessions, 20 min per session)	Lumosity Training 1. Memory Match (Single 2-back, *visual, fixed)* 2. Memory Match overload (Single 3-back, *visual, fixed)* 3. Memory Lane (*Dual N-back, adaptive*)	Active Lumosity Training 1.Processing Speed Speed Match (speeded Single 1-back Task, *visual)* 2.Speed Match overdrive (like Speed Match including partial match option, *visual)* 3.Spatial Speed Match (like the Speed match task but stimuli differ in spatial orientation) II. Passive
Miro-Padilla et al. (64, 65)†	Healthy Adults (N=52, 22.60±1.45) - TG (N=25, 22.77± 1.5) - PCG (N=27, 22.44±1.4)	Randomised controlled trial	3.33 hours, 4 sessions pw (1 week, 4 sessions, 50 min per session)	Single N-back *(letter, adaptive)*	Passive
Opitz et al. (66)	Healthy Adults (N=48, 23.67±2.26*, range = 19-31) - TG (N=16, 23.94±2*, range = 21-29), - ACG (N=16, 23.54±2.4*, range= 20–28), - PCG (N=16, 23.94±2.26*, range= 20–31)	Quasi-experimental	1.5 hours, 4 sessions pw (2 weeks, 9 sessions, 50 min per session)	1.Chinese Vocabulary Learning 2.Single N-back *(visual, adaptive)*	I. Active 1.Chinese Vocabulary Learning 2. Single N-back *(auditory)* II. Passive Chinese Vocabulary Learning, no WMU training
*Roman et al., (68)****	Healthy Adults (N=56, 18.12±1.05)*	Quasi-experimental	12 hours, 2 sessions pw (12 weeks,	Dual N-back *(auditory and visual, adaptive)*	Passive

(Bx Data taken from Colom et al.(91))	- TG (N=28, 18.04±0.9), - PCG (N=28, 18.2±1.2)		24 sessions, 30 min per session)		
Roman et al. (67) *** *(Bx Data taken from Colom et al.(91))*	Healthy Adults (N=56, 18.12±1.05)* - TG (N=28, 18.04±0.9), - PCG (N=28, 18.2±1.2)	Quasi-experimental	12 hours, 2 sessions pw (12 weeks, 24 sessions, 30 min per session)	Dual N-back *(auditory and visual, adaptive)*	Passive
Salminen et al.(69)	Healthy Adults (N=54, 24.5±3.67)* - TG (N=18, 24.4±4), - ACG (N=18, 24.1±3.1), - PCG (N=18, 25±4.0)	Quasi-experimental (no info on randomization)	8 hours, 5 sessions pw (3 weeks, 16 sessions, 30 min per session)	Dual N-back (*auditory & visual, adaptive)*	I. Active Single N-back (*auditory & visual* at different sessions) II. Passive
Schneiders et al. (70)	Healthy Adults (N= 48, 23.67± range= 19-31) - TG1 (N= 16, 23.94± 2.4*, range=21-29), - TG2 (N=16, 23.13±2*, range 20-28), - PCG (N=16, 3.94±2.75*, age range 20-31)	Quasi-experimental	7.5 hours, 4 sessions pw (2 weeks, 9 sessions, 50 min per session)	TG1: Single N-back *(visual)* TG2: Single N-back *(auditory)* *(all adaptive)*	Passive
Schneiders et al. (80)	Healthy Adults (N=32 21.31±1.27*, range=18-24) - TG (N=16, 21.13±1.5, range=18–14), - PCG (N=16, 21.50±1*, range = 19–23)	Quasi-experimental	6.66 hours, 4 sessions pw (2 weeks, 8 sessions, 50 min per session)	Single N-back *(auditory, adaptive)*	Passive
Schweizer et al. (71)	Healthy Adults (N=34, 23± 2.4)	Randomised controlled trial	8.33 hours, 5 sessions pw (4	Dual N-back *(affective, adaptive)*	Active (Feature Matching)

	- TG (N=17, missing data, - ACG (N=15, missing data)		weeks, 20 sessions, 25 min per session)		
Thompson et al. (79) (Bx data taken from (92)	Healthy Adults (N=58, 21.86±2.69) * - TG (N=20, 21.3±2.3) - ACG (N= 19, 21.2±2.0) - PCG (N=19,23.1± 3.3)	Quasi-experimental	13.33 hours, 5 sessions pw (4 weeks, 20 sessions, 40 min per session)	Dual N-back *(auditory & visual, adaptive)*	I. Active (Multiple Object Tracking) II. Passive
Neurological Populations					
Aguirre et al. (51)	Adults with MS (N=29, 32.72±7.48)* - TG (N=15, 35.80 ±7.3), - PCG(N=15, 36.14 ±5.97)	Randomised controlled trial	4 hours, 4 sessions pw (1 week, 4 sessions, 60 min per session)	Single N-back *(verbal, adaptive)*	Passive
Bonzano et al. (55)	Adults with MS (N = 18, 45.3± 10.2)	Quasi-experimental Single group	20 hours, 5 sessions pw (40 Sessions, 30 min per session	1. Dual N-back *(numerical and spatial)* 2. Single N-back (*visuospatial*) 3. Operation N-back *All adaptive*	No CG
Leung et al. (75)	Stroke Participant (N=1, Age = 39 years)	Case study	11.6 hours, 5 sessions pw (7 weeks, 35 sessions, 20 min per session)	Single N-back *(auditory, increased difficulty be default but non adaptive to performance)*	No CG
Leung et al. (76)	Stroke participants (N= 2, Age = 37 years)	Case study	20 hours, 5 sessions pw (6 weeks, 30 sessions, 40 min per session)	Single N-back *(auditory, increased difficulty be default but non adaptive to performance)*	No CG

* The means and SDs to combine groups were calculated based on the formulae provided by Higgins and Deeks (41), p. 177. When the range was reported for individual studies instead of the SD value, then an SD estimate was calculated as the quarter of the range (41), p. 176. In two

cases the SD was either missing or could not be calculated based on other measures of dispersion (63, 71). If studies included more than one control group in their design, then the data were collapsed across them. ****** These studies shared the same dataset, *** These studies shared the same dataset,**** These studies shared the same dataset. **†**These studies share the same dataset; the neuroimaging data on the training effect are described in Miro-Padilla et al. (65) and the neuroimaging data on the transfer effect are described in Miro-Padilla et al. (64) **ACG:** Active Control Group, **Bx:** Behavioural, **DAT-AR:** Differential Aptitude Test – abstract reasoning, **DAT-NR:** Differential Aptitude Test – numerical reasoning subtest, **DAT-VR:** Differential Aptitude Test – verbal reasoning subtest, **EF:** Executive Function, **eWM:** emotional working memory, **HVLT:** Hopkins Verbal Learning Test, **MS:** Multiple Sclerosis, **NA:** Non-Adaptive**, PMA-R**: Primary Mental Abilities – Inductive reasoning subtest, **PMA-V:** Primary Mental Abilities – Vocabulary subtest, **PCG:** Passive Control Group, **RAPM:** Raven's Advanced Progressive Matrices, **STM:** Short Term Memory, **TG:** Training Group, **TG1:** Training Group 1, **TG2:** Training Group 2, **WM:** working memory, **WMU**: WM updating.

Table 2: Neuroimaging protocol details for reviewed studies

Reference	***Neuroimaging Method No of Sites***	***No of scanning sessions***	***Neuroimaging Outcome (Analysis, software)***	***Cognitive Outcome***	
				Changes in performance Criterion Task *(modality)*	***Changes in performance Transfer Task,*** *near or far transfer (modality)*
Healthy Adults					
Aguirre et al. (51)	3T MRI 1	3	Changes in BOLD activity *(task-based fMRI, Whole-brain, SPM12)*	Single N-back *(numerical)*	-
Backman et al. (52)	PET 1	2	Changes in raclopride binding to striatal D2 receptors *(PET, N/A, SPM8)*	Letter memory updating	-
Backman et al. (53)	PET 1	2	Changes in raclopride binding to striatal D2 receptors *(PET, N/A, SPM8)*	Letter memory updating	N-back task, *near (numerical)*
Biel et al. (54)	3T, MRI 1	2	Changes in Grey Matter Volume *(VBM, VBQ, N/A, SPM12)*	-	-
Buschkuehl et al. (77)	No info, fMRI 1	2	Changes in Cerebral Perfusion *(ASL, N/A, MCFLIRT)*	Single N-back *(visuospatial)*	-
*Clark et al. (73)***	3T fMRI 1	2	Changes in BOLD activity *(task-based Fmri, Whole-brain, FSL 5.09)*	Dual N-back *(visual & auditory)*	1.Raven's Standard Progressive Matrices, *far* 2.Lexical Decision, *far*
Colom et al. (56) ***	3T MRI 1	2	Changes in Jacobian determinants *(TBM, N/A, SPM5)*	-	-
*Colom et al. (57)****	3T MRI 1	2	Changes in Grey Matter Volume *(VBM, N/A, SPM8)*	-	-

Dahlin et al. (37)	1.5 fMRI 1	2	Changes in BOLD activity *(task-based fMRI, Whole-brain, SPM2)*	Letter Memory updating (*verbal)*	1.Single N-back, *near (numerical)* 2.Stroop, *far*
Emch et al. (58)	3T, MRI 1	2	Changes in BOLD activity *(task-based fMRI, whole-brain, SPM12)*	Single N-back (*verbal)*	-
Finc et al. (59)	3T, MRI 1	4	Changes functional modularity (*task-based fMRI*, *ROI, fMRIPrep, Nipype*)	Dual N-back *(audio & visual)*	-
Flegal et al. (78)	3T fMRI 1	2	Changes in BOLD activity *(task-based fMRI, ROI, SPM8)*	Matrix Updating *(visuospatial)*	1.Single N-back, *near (visuospatial)* 2.Object Location – Episodic Memory, *far*
*Heinzel et al. (60)*****	3T fMRI, MRI 2	2	Changes in BOLD activity, Functional Connectivity and Grey Matter Volume, *(task-based fMRI & VBM, ROI, SPM8)*	Single N-back *(numerical)*	-
*Heinzel et al. (61) *****	3T fMRI 2	2	Changes in BOLD activity, *(task-based fMRI, Whole-brain & ROI, SPM8)*	Single N-back *(numerical)*	DMS – Maintain & Update Condition, *near*
*Heinzel et al. (62)*****	3T fMRI 2	2	Changes in BOLD activity, *(task-based fMRI, ROI, SPM8)*	Single N-back *(numerical)*	-
Hempel et al.(63) 2004	1.5T fMRI 1	3	Changes in BOLD activity, *(task-based fMRI, VOI, SPM99)*	Single N-back *(spatial)*	-
Kuhn et al. (72)	3T fMRI 1	3	Changes in BOLD activity *(task-based fMRI, Whole-Brain & ROI, SPM5)*	Number Memory Updating *(numerical)*	-
*Lawlor-Savage et al. (74)***	3T MRI 1	2	Changes in GM Surface Area, Thickness & Volume, *(MRI surface-based analysis, N/A, FSL, Freesurfer 5.3.0)*	-	-
Miro-Padilla et al. (64, 65)†	1.5T fMRI 1	3	Changes in BOLD activity	Single N-back *(letter)* †	PASAT, *far* †

			(task-based fMRI, Whole-brain, SPM12)		
Opitz et al. (66)	1.5 fMRI 1	2	Changes in BOLD activity *(task-based fMRI, VOI, Brain Voyager QX)*	-	Chinese Orthographic Task, *far*
*Roman et al. (50)****	3T MRI 1	2	Changes in Cortical Thickness & Surface Area, *(surface-based Morphometry, N/A (FSL, FMRIB Diffusion toolbox, FDT)*	-	-
*Roman et al. (67)****	No info, MRI 1	2	Changes in Structural Connectivity & Fractional Anisotropy, *(DWI, N/A (CIVET pipeline 2.0)*	-	-
Salminen et al. (69)	3T fMRI 1	2	Changes in BOLD activity, *(task-based fMRI, Whole brain & ROI, SPM8)*	1.Dual N-back *(audio and visual)* *2.*Single N-back	1.Dual Letter Memory, *near* 2.Single Letter Memory, *near**,** (verbal)*
Schneiders et al. (70)	1.5T fMRI 1	2	Changes in BOLD activity, *(task-based fMRI,* VOI, Brain voyager QX*)*	-	Visual 2-back: - *near* for Visual Training Group - *far* for Auditory Training Group
Schneiders et al. (80)	3T fMRI 1	2	Changes in BOLD activity, *(task-based fMRI, Whole-brain & ROI, Brain voyager QX)*	-	- Single N-back, *near (auditory)* - Single N-back, *far (visual)*
Schweizer et al. (71)	3T fMRI 1	2	Changes in BOLD activity, *(task-based fMRI, ROI, SPM5)*	Dual N-back *(affective)*	Emotion Regulation task, *far*
Thompson et al. (79)	3T fMRI 1	2	Changes in BOLD activity & Functional Connectivity, *(task-based fMRI, ROI (FSL, Freesurfer)*	Dual N-back *(auditory and visual)*	-
Neurological Populations					

Aguirre et al. (51)	3T MRI 1	3	Changes in BOLD activity *(task-based fMRI, Whole-brain, SPM12)*	Single N-back *(numerical)*	-
Bonzano et al. (55)	1.5 MRI 1	2	Changes in BOLD activity *(task-based fMRI, Whole-brain, SPM12)*	-	PVSAT, *far*
Leung et al. (75)	1.5T fMRI 1	2	Changes in BOLD activity *(task-based fMRI, Whole-brain, SPM8)*	Single N-back*(auditory)*	-
Leung et al. (76)	1.5T fMRI 1	2	Changes in BOLD activity *(task-based fMRI, Whole-brain, SPM8)*	Single N-back*(auditory)*	-

These studies shared the same Bx dataset., * These studies shared the same Bx dataset., **** These studies shared the same Bx dataset. **†**These studies share the same dataset; the neuroimaging data on the training effect are described in Miro-Padilla et al. (65) and the neuroimaging data on the transfer effect are described in Miro-Padilla et al. (64) **ASL:** Arterial Spin Labelling, **Bx:** Behavioural, **D2:** Dopamine 2, **DMS**: Delayed Match-to-Sample **DWI:** Diffusion Weighted Imaging, **fMRI:** functional Magnetic Resonance Imaging, **FSL:** FMRIB Software Library, **GM:** Gray Matter, **HAWIE-R**: Hamburg–Wechsler Adult Intelligence Scale–Revised, **N/A:** Non-Applicable, **Nipype:** Neuroimaging in Python Pipelines and Interfaces, **PASAT:** Paced Auditory Serial Addition Test, **PVSAT:** Paced Visual Serial Addition Test, **ROI:** Region of Interest, **SPM:** Statistical Parametric Mapping, **TBM:** Tensor Based Morphometry, **VBM:** Voxel-Based Morphometry, **VBQ:** Voxel-Based Quantification, **VOI:** Volume of Interest.

Table 3: Cognitive Performance Changes after WMU training for training and transfer tasks

Reference	***Training***	***Transfer***
Healthy Adults		
Aguirre et al. (51)≠	*Training (both Healthy and Multiple Sclerosis training groups collapsed) by Time comparison* 2-back accuracy: non significant. 3-back accuracy: Both TGs significantly improved accuracy on the 3-back level after training compared to the CGs, ($F(2,51) = 10.18$, $p<0.001$, $\eta^2 = 0.29$). *Training (TG vs CG) by Group (Healthy vs Multiple Sclerosis) comparison, ns* *Training by Group by Session, ns*	-
Backman et al. (52)	Updating training significantly improved letter-memory performance ($p<0.001$, $d=1.7$). The training group ($p < 0.001$), but not the controls ($p> 0.20$) improved after training.	Near transfer (n-back task)‡ significant transfer effect ($P < 0.01$, $d = 0.98$)
Backman et al. (53)	*Session by Group comparison* TG showed larger performance gains after training than the CG, ($F (1, 23) = 24.579$, $p < 0.001$, η^2partial = 0.52; $d = 2.07$)	*Session by Group comparison* There were no time effects as a function of group ($p > 0.05$; $d = 0.00$). No behavioural transfer effects were observed.
Biel et al. (54)	Both TG1 and TG2 improved their performance over time‡ *Main effect of time* - Correct Hit Rates (cHR): $F(1,53) = 227.293$, $p < .001$, partial $\eta^2 = .811$, higher cHR post-test - Reaction Times (RTs): $F(1,53) = 51.830$, $p < .001$, partial $\eta^2 = .494$, faster RTs post-test	*Session by group comparison* ‡ - *Far transfer, Processing speed* (d2-R working speed: $F(2,80) = 3.588$, $p =0.032$, partial $\eta^2 = .082$) - *Near transfer, Verbal memory* (VLMT learning: $F(2,80)= 3.254$, $p=0.044$, partial $\eta^2 = .075$). Comparisons did not survive Bonferroni corrections so there is no evidence of transfer.
Buschkuehl et al. (77)	*Session by group by load comparison*	Near transfer (auditory n-back task)‡ *Session by Group comparison* ($p<0.001$)

	TG improved more from pre to post than the CG in the 4-back load condition, ($F(1,52)=12.41$, $p<.001$, η^2partial $=0.19$).	TG improved on the auditory n-back task compared to the CG, specifically driven by the 3-back condition.
*Clark et al. (73)***	*Group by time comparison* Better performance in the TG compared to the CG for the 3-back condition ($F(1,47)=17.04$, $p < 0.001$).	*Far transfer* *Group by time comparison* - *RSPM transfer task* Better performance in the TG compared to the CG for the hardest difficulty level ($F(1,47)=5.88$, $p = 0.019$). This effect was driven by worst performance in the CG post training. - *Lexical decision task* Better performance in the TG compared to the CG for the easy condition ($F(1,47)=5.37$, $p = 0.019$). This effect was driven by significantly different performance before training ($t_{47}=2.0$, $p=0.043$).
*Colom et al. (56)****	Participants improved for both single and dual versions of the training tasks. They engaged in the training protocol and reached the required performance levels by the end of the training. For the visual condition the improvement was 41%, for the auditory condition it was 39%, and for the dual condition it was 53% across the training sessions as reported in Colom et al. (91) ‡ There were large individual differences in the level achieved ranging from 3- to 9-back.	-
*Colom et al. (57)****	*Same as Colom et al. (56)*	-
Dahlin et al. (37)	- Experiment 1 – Young adult Group *Group by session comparison* TG showed larger gains in letter memory compared to the CG, ($F(1,20) =26.45$, $P < 0.001$). - Experiment 2 – Older adult Group *Group by session comparison* TG showed larger gains in letter memory compared to the CG, ($F(1,17) =20.56$, $P < 0.001$).	- Experiment 1 - Young adult Group *Near Transfer* *Group by session comparison* TG performed better compared to the CG post training ($F1.20 = 10.32$, $P < 0.01$) for the 3-back condition, and the effect size for TG was significantly greater than for CG. *Far Transfer*

		No significant training-related changes in performance - Experiment 2 – Older adult Group *Near Transfer* No significant training-related changes in performance
Emch et al. (58)	*Group by Session Comparison for 3-back load,* ($F(1,55)=18.07$, $p<0.001$). Post hoc analyses revealed no significant improvement in the CG ($p = 0.06$), but a highly significant improvement in the TG ($p< 0.001$).	*Near transfer* (HAWIE-R forward & backward) ‡ - *HAWIE-R forward* Group by Session Comparison, ($F(1,55) = 17.248$, $p < 0.001$). Post hoc analyses revealed a performance decrease in the CG ($p = 0.045$) and a highly significant improvement in the TG ($p < 0.001$). - *HAWIE-R backward* Group by Session Comparison, ns.
Finc et al.(59)	*Session by Condition by Group comparison (*$\chi^2(3)= 9.39$, $p= 0.02$) - *2-back:* TG exhibited significantly larger training gains post-training compared to the CG ($t(20)= -4.12$, $p=0.004$) - *1-back:* t-test comparison between groups, ns ($t(39.64) = -0.52$, $p = 0.47$)	-
Flegal et al. (78)	*ANCOVA on post-training performance, controlling for pre-training performance* TG improved performance compared to the CG for the 7-updates condition $F(2,52)=4.50$, $p < .05$, η^2partial=0.15)	*ANCOVA on post-training performance, controlling for pre-training performance* - *Near Transfer* No significant differences between the groups. - *Far Transfer* TG improved performance compared to the CG for the 8-associates condition $F(2,52)=4.50$, $p < .05$, η^2partial=0.15)
Heinzel et al. (60)	The older adult TG improved overall after training for all three difficulty levels, 1-,2- and 3-back while the strongest improvement was found for the 2-back condition.	-
Heinzel et al. (61)	*Group by Time comparison*	*Near transfer* (DMS task, maintain and update conditions)

	The TG showed stronger improvement in the n-back task compared to the CG in accuracy (F (1,27)=24.07, p<0 .001, η² partial =0.47) and reaction times (F(1,27) = 11.22, p =0.002, η² partial= 0.29).	*Group by Time comparison* The TG showed stronger improvement for the maintain 5 condition only compared to the CG (F(1,27)=4.92, p=0.035)
Heinzel et al. (62)	*Group by Time by Load comparison* The TG improved more compared to the CG in the 1-back, 2-back and 3-back load conditions.	*Near transfer (DMS* (single and dual versions, auditory and visual) ‡ - *Visual Single Task* General improvement in task performance but no significant differences between TG and CG post-training (Group by Time, ns; Main effect of group, ns; Main effect of Time, p=0.025). - *Auditory Single-Task* No significant differences between TG and CG post-training as well as no evidence for performance improvement over time. (Group by time, ns, Main effect of time, ns; Main effect of Group, ns). - *Dual Task (Accuracy)* Group by Time comparison, p=0.038 Training-related improvement for the TG compared to the CG. - *Transfer Dual Task* Absolute performance (%correct), Group by Time comparison TG improved dual-task performance compared to the CG Relative performance, Group by time by load by modality comparison The dual-task costs decreased in the TG compared to the CG for the auditory modality post-training in the 1-load condition.
Hempel et al. (63)	The mean rate of relative errors improved significantly for the 2-back condition between the first and second sessions and remained stable in the third session. No significant changes for the 0-back and 1-back conditions.	-

Kuhn et al. (72)	*Group by Time by load comparison* TG improved more compared to the CG post-training especially for the higher load condition.	*Near transfer (numerical N-back and spatial updating)* ‡ *Group by Time comparison* Significant linear and quadratic effects of time for both *n-back* and *spatial updating* tasks., ps<0.04. Only non-significant trends favouring the TG compared to the CG.
*Lawlor-Savage et al. (74)***	*TG* Correct matches significantly increased in all training tasks (comparison of the average of the first five iterations of each game to the last five iterations of each game) ‡ *CG* Reaction times significantly decreased in all three training tasks (comparisons of the average of the first five to the last five games). ‡	-
Miro-Padilla et al. (64, 65) †	*Group by time by load comparison* TG performed better than CG in both sessions post-training in both accuracy and reaction times measures in the 2- and 3-back load levels.	*Far Transfer (PASAT)* *Group by Session comparison, ns.* TG did not perform the task significantly better than the CG after *n*-back training, no evidence of transfer.
Opitz et al. (66)	*Time by Group comparison, ns* The visual TG as well as the active auditory CG improved their performance in the course of training as revealed by a significant main effect of session, ($F_{7,24}=11.58$, $p<0.001$, $\eta^2_{partial}=0.77$).	Far transfer (Chinese orthographic task) *Time by Group comparison, ns* Performance increased significantly from pre- to post-test only for visual TG [mean difference =.08, SD=0.13, $t_{15}=-2.68$, $p<0.05$] but not for the ACG and PCG.
Roman et al. (68) ***	*Same as Colom et al. (56)*	-
Roman et al. (67) ***	*Same as Colom et al. (56)*	-
Salminen et al. (69)	*Group by Session comparison* *Dual n-back* TG shows greater improvement compared to the ACG and PCG, $F_{(2,50)}=25.06$, $p<0.001$, $\eta^2_{partial}=0.50$) *Single n-back*	*Near transfer (dual and single WMU task)* *Group by Session comparison* - *Dual WM updating task* TG shows improved performance following training while the ACG and PCG showed no changes in performance.

	TG and ACG show equal improvement, F (2,51)=15.40, p<0.001, η²partial=0.38).	- *Single WM updating task* No significant interaction, all groups showed improved performance in both auditory and visual versions.
Schneiders et al. (70)	-	*Visual n-back task* *Group by time comparison, F=2,45=3.52, p<0.05, ηp2=0.14* Group specific performance improvements. - *Near transfer for visual TG* The Visual TG significantly improved after training, [F(1,15)=36.01, p<0.001, η²partial=0.71]. - *Far transfer for auditory TG* *Auditory TG* The auditory TG didn't exhibit significant improvement post training, F(1,15)=3.73, p<0.10, η²partial=0.20, ns.
Schneiders et al. (80)	-	*Auditory transfer task (near)* *Time by Group comparison,* [F(1, 30) = 25.23, p < 0.001, η2p= 0.46] Post-test performance was significantly greater in the TG compared to no training (t(30) = 4.23, p < 0.001). *Visual transfer task (far)* No significant differences in the groups post-training.
Schweizer et al. (71)	*Time by Group comparison* Significant pre to post training increase in performance for the TG while for the CG, no changes were evident.	*Far transfer (Emotion Regulation task)* *Time by Group comparison* The TG exhibited significantly greater reduction in emotional distress to negative films in the Regulate relative to the Attend condition compared to the CG. TG showed a decrease in emotional distress post-training (Regulate relative to attend condition) while the CG exhibited a non-significant
Thompson et al. (79)	*Session by Group comparison* TG improved more after training compared to the CG specifically for the highest load conditions, i.e. 2- and 3-back in both accuracy and reaction times.	-

Neurological Populations		
Aguirre et al.(51) ≠	*Training (both Healthy and Multiple Sclerosis training groups collapsed) by Time comparison* *2-back:* ns 3-back: Both TGs significantly improved accuracy on the 3-back level after training compared to the CGs, ($F(2,51) = 10.18$, $p<0.001$, $\eta^2 =0.29$). *Training (TG vs CG) by Group (Healthy vs Multiple Sclerosis) comparison, ns* *Training by Group by Session, ns*	-
Bonzano et al.(55)	-	*Rao's Brief Repeatable Battery of Neuropsychological Tests (BRB-NT)* ‡ MS Patients improved significantly post-test in all BRB-NT subtests (all $p_s< 0.05$). *PVSAT* Performance data in this task is not reported.
Leung et al. (75)	An average Cohen's d value of 4.11 for the pre-training and post-training assessments indicating a better than chance performance (1-back & 2-back conditions only).	-
Leung et al. (76)	*Participant 1* exhibited longer reaction times for both 1-back and 2-back conditions in the post compared to the pre-training sessions, while the hit rate improved. *Participant 2* showed improvement in both hit rate and reaction times for the 1-back and 2-back conditions post-training.	-

The group comparison tests and p values in this table were extracted directly from each study as reported by the authors. ≠ The data are collapsed across both training groups (Healthy adults and patients with MS); the F and p values cannot be reported for each group separately.

** These studies shared the same Bx dataset., *** These studies shared the same Bx dataset, †These studies share the same dataset; the neuroimaging data on the training effect are described in Miro-Padilla et al. (65) and the neuroimaging data on the transfer effect are described in Miro-Padilla et al. (64), ‡This data refers to cognitive tasks that were assessed outside of the scanner, **Bx:** Behavioural, **MS:** Multiple Sclerosis, **ns:** non significant, **PASAT:** Paced Auditory Serial Addition Test, **PVSAT:** Paced Visual Serial Addition Test, **TG1:**Training Group1, **TG2:** Training Group 2, **VLMT:** Verbal Learning Memory Test. The group comparison tests and p values in this table were extracted directly from each study as reported by the authors.

Table 4: Functional Activity Changes after WMU training for training and transfer tasks

Reference	***Functional Activity Changes***	
	Training	***Transfer***
Healthy Adults		
Aguirre et al. (51) ≠	*Training (both Healthy and Multiple Sclerosis training groups collapsed) by Time comparison,* p<0.05 FEW corrected, p <0.001 and uncorrected. **2-back: ↓*Decreased* activity TG vs CG in:** i. R Angular gyrus ii. R Supramarginal gyrus iii. L/R Inferior parietal lobule iv. R middle frontal gyrus v. L Postcentral gyrus **3-back: ↓*Decreased* activity TG vs CG in:** i. R Superior medial frontal gyrus ii. L/R Middle frontal gyrus iii. L/R Superior frontal gyrus iv. L/R Supplementary motor area v. L Precentral gyrus vi. L Inferior frontal gyrus	-
Backman et al. (52)	*Group by Time comparison, threshold at p<0.001* ↓ ***Decreased raclopride binding to D2 receptors*** for the TG compared to the CG in the L caudate. Enhanced DA release after cognitive training is demonstrated. Suggestive of ↑ ***Increase in caudate BOLD activity.***	-
Backman et al. (53)	*Group by Time comparison* ↓ ***Decreased raclopride binding to D2 receptors*** for the TG compared to the CG in L/R Striatum. Enhanced DA release after cognitive training is demonstrated Suggestive of ↑ ***Increase in caudate BOLD activity.***	Group by Time comparison ↓ ***Decreased raclopride binding to D2 receptors*** for the TG compared to the CG in R Striatum.

Buschkuehl et al. (77)	Group by Time comparison (4-back Vs 1-back) (threshold: z > 2.8; cluster size >= 19) ↑ ***Increase in magnitude of perfusion*** *for TG compared to the CG in:* i. R Frontal postcentral gyrus ii. L Superior frontal gyrus (BA6) iii. R superior occipital gyrus iv. R middle occipital gyrus Group by Time comparison (4-back Vs 1-back) ↑ ***Increase in perfusion changes at rest*** *for TG compared to the CG in:* i. L Frontal precentral gyrus (BA6) ii. L Parietal Angular Gyrus (BA39) ↑ ***Decrease in perfusion changes at rest*** *for TG compared to the CG in* R postcentral gyrus (BA5).	-
Clark et al.(73)	*Group by Time comparison* Z threshold of 2.3 and cluster threshold of 0.05 ↓ ***Decreased activity post-training*** for the TG compared to the ACG: i. L/R paracingulate gyrus ii. L/R anterior cingulate gyrus iii. L/R frontal pole iv. L/R superior frontal gyrus v. L/R cingulate gyrus vi. L/R insular cortex vii. L/R temporal pole viii. L/R parahippocampal gyrus ix. L/R posterior cingulate gyrus x. R middle temporal gyrus R angular gyrus xi. R supramarginal gyrus xii. L/R posterior cingulate gyrus xiii. L postcentral gyrus Brain regions combined from the following contrasts: 3-back > 2-back, 3-back > 1-back and 2-back > 1-back.	***Transfer (far)*** Group by Time comparison ↑ ***Increased activity post-training*** for the TG compared to the ACG: i. L Inferior Frontal gyrus; ii. L Frontal pole iii. L Precentral gyrus; iv. L Postcentral gyrus; v. L Superior Frontal gyrus **Hard > Medium Condition**

Dahlin et al. (37)	**Experiment 1: Young Adult Group** *Group by Session comparison* ↑ ***Increased activity*** for the TG post-training in: i. L/R striatum ii. R Temporal lobe iii. R Occipital lobe ↓ ***Decreased activity*** for the TG post-training in: i. L Frontal lobe ii. L Parietal lobe **Experiment 2: Older Adult Group** *Group by Session comparison* ↑ ***Increased activity*** for the TG post-training in: i. L Frontal lobe ii. L/R Parietal lobe iii. R Temporal lobe iv. L Cerebellum v. L Striatum	**Experiment 1: Young Adult Group** ***Transfer (near)*** Group by Session comparison ↑ ***Increased activity*** for the TG post-training in: i. L Frontal lobe ii. L Parietal lobe iii. L Temporal lobe iv. L Striatum v. Brain stem ***Transfer (far)*** No changes. **Experiment 2: Older Adult Group** No significant changes were found for the 3-back task
Emch et al. (58)	*Group by Time comparison FDR corrected p<0.05, k=6 voxels.* ↓***Decreased activity for TG compared to CG post-training in:*** i. L middle temporal gyrus (BA20, BA39) ii. R superior frontal gyrus (BA9) iii. L/R supramarginal gyrus (BA40) iv. R anterior cingulate (BA32) v. R posterior cingulate (BA29) vi. L cuneus (BA7) vii. R middle frontal gyrus (BA9) viii. R angular gyrus (BA39) ix. R middle occipital gyrus (BA19) x. R occipital lobe (BA18) xi. L parahippocampal gyrus (BA30) xii. L/R cerebellum	-
Flegal et al. (78)	*Group x Time comparison all clusters above p<0.05.*	*Group by Time comparison*

	Matrix Updating task ↓***Decreased activity*** greater for the TG compared to CG post-training in all ROIs: i. L/R Caudate ii. L/R Putamen iii. L/R Hippocampus **Whole brain analysis** cluster corrected FWE threshold, p <.05. *Group by session interaction* ↓***Decreased activity for TG:*** xiii. L/R striatum, xiv. L/R prefrontal, xv. L/R temporal, xvi. L parietal regions xvii. L parietal regions.	No significant differences between TG and CG for the near transfer and far transfer tasks in the ROIs.
Heinzel et al. (60)	*Time by Load comparison* p<0.05 FWE corrected for whole brain). ↓***Decreased activity*** for the TG post-training in the WM network: i. L/R Rostral Cingulate Zone (BA32/6) ii. L/R lateral premotor cortex (BA6) iii. L/R DLPFC (BA9/46) iv. L/R Intraparietal sulcus (BA40) Follow-up t-tests indicating the effect was driven by 1-back load.	-
Heinzel et al. (61)	*Group x Time comparison*, all clusters above p<0.05. **Combined 1&2-back (k>90, alphasim-corr)** ↓ ***Decreased activity*** for the TG post-training in: i. R/L Medial Frontal gyrus / Anterior Cingulate gyrus/ Supplementary Motor area (k=166) ii. R Middle and Superior Frontal gyrus (k=140) R supramarginal gyrus, Inferior Parietal lobule, and angular gyrus (k=112)	***Transfer (near)*** Group x Time comparison, all clusters above p<0.05. Sternberg Updating 3&5 **(k>57, alphasim-corr)** ↓ ***Decreased activity*** for the TG post-training in the R middle frontal gyrus/superior frontal gyrus (k=68)
Heinzel et al.(62)	Same as in Heinzel et al., 2014.	-

Hempel et al. (63)	Changes in mean effect sizes for 2-back and 1-back load levels for the TG. k=20 voxels; p<0.05, corrected for multiple comparisons) Significant Inverse U-Shape Quadratic Function for the mean effect size: ↑ ***Increased activity*** between sessions 1 and 2 in R Intraparietal sulcus/ superior parietal lobe for both 1 and 2-back load levels. ↓ ***Decreased activity*** between sessions 2 and 3 in R Intraparietal sulcus/ superior parietal lobe Non-significant quadratic trend for the mean effect size in the R inferior/medial frontal gyrus.	-
Kuhn et al.(72)	Contrast of all load conditions against implicit baseline averaged over group and time point. (threshold p<0.01, cluster>22) ↑ ***Increased activity*** for the TG between sessions 1 and 2 in R/L striatum (putamen). ↓ ***Decreased activity*** for the TG between sessions 2 and 3 in: i. R striatum (putamen) ii. R inferior frontal gyrus	-
Miro-Padilla et al. (64, 65) †	*Group by Session* comparison separately for each load level (2-back and 3-back). p < 0.05 FWE cluster-corrected using a threshold of p < 0.001 at the uncorrected voxel level ↓ ***Decreased activity*** for the TG between sessions 1 and 2 in: i. R Frontal Superior (BA32/6) ii. L Frontal Middle (BA10) iii. R Frontal Middle (BA6) iv. L Parietal Inferior (BA40) v. R Parietal Inferior (BA40) vi. L Temporal Middle (BA21) vii. L Frontal Superior (BA6)	***Transfer (far),*** *Group by session comparison, FDR threshold of p< 0.05* ↓***Decreased activity*** for the TG compared to the CG post-training in the R Dorsolateral Prefrontal Cortex (BA 46).

	viii. R Frontal Middle (BA46) ix. R Parietal Inferior (BA40) x. SMA (BA6) xi. L Frontal Inferior (BA48) Brain regions combined from 3-back and 2-back levels.	
Opitz et al. (66)	-	***Transfer (far)*** Time by Group comparison **PCG** ↓ ***Decreased activity*** in L fusiform gyrus. No significant changes for either of the TG or ACG.
Salminen et al. (69)	AlphaSim correction p<0.001, cluster size>22. ↓ ***Decreased activity*** for the TG post-training in: i. R Inferior frontal gyrus ii. R Middle frontal gyrus iii. R Superior frontal gyrus iv. L Medial frontal gyrus v. L Superior frontal gyrus vi. R Inferior Parietal lobule vii. R Anterior cingulate gyrus viii. L Posterior cingulate gyrus ix. R Cerebellum x. L Cerebellum ↑ ***Increased activity*** for the TG post-training in L precentral gyrus. ↓ ***Decreased activity*** for the ACG post-training in: i. R Middle Frontal gyrus ii. L Inferior Frontal gyrus iii. L Inferior Parietal lobule ***No training-related activation changes for the PCG.***	***Transfer (near)*** ↑ ***Increased activity*** for the TG post-training in: i. L/R calcarine gyrus, cuneus ii. L/R Striatum ***No activation changes for the ACG or PCG for the transfer task.*** Group by Time comparison, for percentage signal changes (PSC), (AlphaSim p <0.001, cluster size >22) ↑ ***Increased activity*** for the TG pre to post-training in the striatum ↓ ***Decreased activity*** for the ACG & PCG pre to post-training in the striatum
Schneiders et al. (70)	-	***Transfer (near)*** *Group by Time comparison* ↓ ***Decreased activity*** for the visual TG post-training in: i. R middle frontal gyrus (BA9) ii. R middle frontal gyrus (BA/46) ***Transfer (far)***

		No significant changes post-training for the auditory TG. **Training (across modal training effects)** Group (collapsed across TGs vs CG) by Time comparison, ↓ ***Decreased activity*** for both TGs compared to the CG post training in: i. R Intraparietal sulcus (BA40) ii. R Superior Middle frontal gyrus
Schneiders et al. (80)	-	***Transfer (near), auditory task &*** Group by Time comparison Percent signal change values of functional volumes of interests thresholded at p <0.005 (135 voxel extend ↓ ***Decreased activity*** for the TG post-training in: iii. R Inferior frontal gyrus (BA46) iv. R Inferior frontal gyrus (BA47) Decreased activity was larger in the near transfer task compared to the far transfer task. ***Transfer (far), visual task*** No significant group by time comparison
Schweizer et al. (71)	Group by Time comparison, FDR, p<0.05 ↓ ***Decreased activity*** for the TG compared to the CG post-training across all n-back levels in: i. L ventrolateral to dorsolateral prefrontal cortex ii. L/R inferior parietal cortex iii. R precuneus iv. Inferior/middle temporal gyrus v. L/R middle and posterior cingulum vi. L ACC	**Transfer (far, ER task)** Whole-brain level, p uncorrected<0.001, Regulate relative to Attend condition, TG Vs CG ↑**Increased activity** for the TG compared to the CG post training in: R superior temporal gyrus
Thompson et al. (79)	*Time by Group comparison* ↓ ***Decreased activity*** for the TG compared to the CG post-training in: i. Prefrontal cortex ii. Parietal cortex iii. Insular cortex	-

Neurological Populations		
Aguirre et al. (51) ≠	*Training (both Healthy and Multiple Sclerosis training groups collapsed) by Time comparison,* p<0.05 FWE corrected, and p <0.001 uncorrected. **2-back: ↓*Decreased* activity TG vs CG in:** i. R Angular gyrus ii. R Supramarginal gyrus iii. L/R Inferior parietal lobule iv. R middle frontal gyrus v. L Postcentral gyrus **3-back**: **↓*Decreased* activity TG vs CG in:** *i.* R Superior medial frontal gyrus *ii.* L/R Middle frontal gyrus *iii.* L/R Superior frontal gyrus *iv.* L/R Supplementary motor area *v.* L Precentral gyrus *vi.* L Inferior frontal gyrus	-
Bonzano et al.(55)	-	***Transfer (far)*** Paired t-test (p<0.001 uncorrected, k=30 voxels) ↓***Decreased*** activity found post-training compared to pre-training for the TG in: i. L Cingulate gyrus ii. R postcentral gyrus iii. L inferior parietal lobule
Leung et al. (75)	All activations significant at p<0.005, cluster size>196ml **Main effect of time** (Pre> Post-training) ↓ **Decreased activity** in R Angular gyrus.	-
Leung et al. (76)	All activations significant at p<0.005, cluster size>196ml - Participant 1 – 1-back level ↓ **Decreased activity** in R Middle temporal gyrus (BA37) ↑ **Increased activity** in L temporal gyrus (BA20) - Participant 1 – 2-back ↓ **Decreased activity** in	-

	i. R Middle temporal gyrus (BA37) ii. L Inferior parietal lobe (BA40) ↑ **Increased activity** in R Middle temporal gyrus (BA20) - Participant 2- 1-back level ↓ **Decreased activity** after training in: i. R Middle frontal lobe (BA6) ii. R Inferior frontal gyrus (BA45) iii. R Middle temporal gyrus (BA21) iv. L/R Inferior parietal lobe (BA7/BA40) 2-back level ↓ **Decreased activity** after training in L/R Middle frontal gyrus (BA45/47) ↑**Increased activity** after training in: i. L middle temporal gyrus (BA20) ii. L/R inferior parietal lobe (BA40) iii. R Cerebellum	

≠ The data are collapsed across both training groups (Healthy adults and patients with MS); the F and p values cannot be reported for each group separately, **†**These studies share the same dataset; the neuroimaging data on the training effect are described in Miro-Padilla et al. (65) and the neuroimaging data on the transfer effect are described in Miro-Padilla et al. (64), **FDR:** False Discovery Rate, **FWE:** Family-Wise Error.