Working Memory Training Revisited: A Multi-Level Meta-Analysis of N-back Training Studies

Anna Soveri[1], Jan Antfolk[1], Linda C. Karlsson[1], Benny Salo[1], Matti Laine[1,2]

[1]Department of Psychology, Åbo Akademi University, Turku, Finland

[2]Turku Brain and Mind Center, Turku, Finland

Abstract

The efficacy of working memory (WM) training has been a controversial and hotly debated issue during the last years, and despite a large number of training studies and several meta-analyses, the matter has not yet been solved. We conducted a multi-level meta-analysis on the cognitive transfer effects in healthy adults who have been administered WM updating training with n-back tasks, the most common experimental WM training paradigm. Thanks to this methodological approach that has not been employed in previous meta-analyses in this field, we were able to include effect sizes from all relevant tasks used in the original studies. Altogether 203 effect sizes were derived from 33 published randomized controlled trials. In contrast to earlier meta-analyses, we separated task-specific transfer (here untrained n-back tasks) from other WM transfer tasks. Two additional cognitive domains of transfer that we analyzed consisted of fluid intelligence (Gf) and cognitive control tasks. A medium-sized transfer effect was observed to untrained n-back tasks. For other WM tasks, Gf, and cognitive control, the effect sizes were of similar size and very small. Moderator analyses showed no effects of age, training dose, training type (single vs. dual), or WM and Gf transfer task contents (verbal vs. visuospatial). We conclude that a substantial part of transfer following WM updating training with n-back is task-specific and discuss the implications of the results to WM training research.

Keywords: Cognitive training, Working memory, Executive functions, N-back, Meta-analysis

Working Memory Training Revisited: A Multi-Level Meta-Analysis of N-back Training Studies

Working memory (WM) training has stirred considerable interest amongst researchers and public at large during the last decade or so (for reviews, see e.g., von Bastian & Oberauer, 2014; Green & Bavelier, 2008; Klingberg, 2010; Lövdén, Bäckman, Lindenberger, Schaefer, & Schmiedek, 2010; Morrison & Chein, 2011). The main reason for this widespread interest is that WM has been linked to a number of important skills such as academic achievement and general intellectual capacity (for reviews, see Engle, 2002; Shipstead, Redick, & Engle, 2010). Moreover, WM deficits often occur in common clinical conditions such as dyslexia, ADHD, and major depression, as well as in normal aging (for a review, see Lezak, Howieson, & Loring, 2004). As a system for short-term maintenance and manipulation of task-relevant information (Baddeley, 2000), WM is inherently involved in all higher-level cognitive activities. Accordingly, WM training, if successful, might have wide-reaching consequences for an individual.

The results from the initial WM training studies were very promising as they suggested that it is possible to improve performance not only on the trained task, but also on untrained tasks measuring other cognitive functions (Jaeggi, Buschkuehl, Jonides, & Perrig, 2008; Klingberg, Forssberg, & Westerberg, 2002). The initial enthusiasm, however, turned into a controversy as subsequent training studies reported mixed results (e.g., Brehmer, Westerberg, Bäckman, 2012; Bürki, Ludwig, Chicherio, & De Ribaupierre, 2014; Bäckman et al., 2011; Chooi & Thompson, 2012; Colom et al., 2013; Dahlin, Neely, Larsson, Bäckman, & Nyberg, 2008; Jaeggi et al., 2008, 2010; Klingberg et al., 2002, 2005; Kundu, Sutterer, Emrich, & Postle, 2013; Lilienthal, Tamez, Shelton, Myerson, & Hale, 2013; Oelhafen et al., 2013; Redick et al., 2013; Salminen, Strobach, & Schubert, 2012; Thompson et al., 2013; Waris, Soveri, & Laine, 2015). Here the key issue is the existence of generalization following

cognitive training, as the aim is to elicit positive transfer effects on untrained tasks. Most training studies make a distinction between near and far transfer effects. Near transfer refers to enhanced performance in a task that is intended to measure the trained cognitive domain, and far transfer to improvement in another cognitive domain, such as WM training leading to better performance in a task measuring intelligence (e.g., von Bastian & Oberauer, 2014). As WM training studies have reported near transfer, far transfer, both near and far transfer, or no transfer at all, it has been difficult to draw conclusions about the efficacy of WM training. What complicates the matter further is that many of the previous training studies have suffered from methodological shortcomings, such as using small sample sizes, employing a no-contact control group, failing to randomly assign participants to groups, or using only a single task to measure a given cognitive ability (see e.g., Melby-Lervåg & Hulme, 2013; Melby-Lervåg, Redick, & Hulme, 2016; Morrison & Chein, 2011; Shipstead et al., 2010; Shipstead, Redick, & Engle, 2012).

Previous Meta-Analyses on Working Memory Training

Given the widespread interest in WM training and the large variability in results, it is not surprising that during the past five years a number of meta-analyses have addressed the outcomes of WM training (Au et al., 2015; Dougherty, Hamovitz, & Tidwell, 2016; Melby-Lervåg, & Hulme, 2013; Melby-Lervåg & Hulme, 2016; Melby-Lervåg et al., 2016; Schwaighofer, Fischer, & Bühner, 2015; Weicker, Villringer, & Thöne-Otto, 2016; for meta-analyses investigating not only WM training, see Hindin & Zelinski, 2012; Karbach, & Verhaeghen, 2014; Karr, Areshenkoff, Rast, & Garcia-Barrera, 2014; Kelly et al., 2014). Near and far transfer effects of different kinds of WM training have been studied in four meta-analyses (Melby-Lervåg & Hulme, 2013; Melby-Lervåg et al., 2016; Schwaighofer et al., 2015; Weicker et al., 2016). In sum, the results from these meta-analyses (Table 1) with both clinical and healthy samples of children and adults showed that WM training can produce

small to large short-term near transfer effects (verbal and visuospatial WM tasks), and null to small effects of transfer to verbal and visuospatial reasoning (or fluid intelligence; Gf) tasks. The results also showed small transfer effects to cognitive control (Melby-Lervåg & Hulme, 2013; Weicker et al., 2016), and attention (Weicker et al., 2016), but no transfer to long-term memory (Weicker et al., 2016), arithmetic skills (Melby-Lervåg & Hulme, 2013; Schwaighofer et al., 2015; Melby-Lervåg et al., 2016), reading comprehension (Melby-Lervåg et al., 2016), or word decoding (Melby-Lervåg & Hulme, 2013; Schwaighofer et al., 2015; Melby-Lervåg et al., 2016).

As is evident from the short review above, there is some disagreement between the previous meta-analyses, particularly regarding the size of the near transfer effects. These discrepancies are at least partly due to differences in decisions regarding the studies included in the meta-analyses and the categorization of tasks into different cognitive domains. For example, the meta-analysis by Melby-Lervåg et al. (2016) included also studies that were not randomized controlled trials. Furthermore, all previous meta-analyses included only one measure per domain per study. In some cases, this measure was a single task selected to represent a domain in a study, while in others, a mean value across the tasks was used. Moreover, the choice of which tasks to include varies between the meta-analyses. For example, unlike the other meta-analyses, Schwaighofer et al. (2015) did not include the n-back task in the near transfer measures due to validity issues with this task. Also, simple span tasks such as the digit span were considered WM measures in the Melby-Lervåg and Hulme (2013) study, while the other meta-analyses coded simple spans as measures of short-term memory (STM; Schwaighofer et al., 2015; Weicker et al., 2016) or excluded them completely from the analyses (Melby-Lervåg et al., 2016). Finally, a close look at the measures in Melby-Lervåg and Hulme (2013) shows three instances where the trained task itself was included in the near transfer analyses, likely leading to an undue inflation of the corresponding effect size.

Of the previous meta-analyses, Melby-Lervåg and Hulme (2013) reported the strongest effect

sizes for near transfer. Given these issues concerning the previous meta-analyses, further

meta-analyses in this popular research field are called for.

Background for the Present Study

The variety of WM training programs, experimental designs, and participant groups

makes it challenging to draw definite conclusions on the WM training outcomes. To limit

three sources of variability (training program, age, health status), the present meta-analysis

focused solely on studies with healthy adults who trained with the most commonly used

computerized experimental WM training paradigm, namely with n-back tasks. In the n-back

task (Kirchner, 1958), the participant is presented with a sequence of stimuli and the task is to

decide for each stimulus whether it is the same as the one presented *n* trials earlier. In a single

n-back task, the participant is required to attend to one stream of stimuli and in a dual n-back

task, two streams of stimuli are presented simultaneously. Figure 1 shows an example of a

dual n-back task with an auditory-verbal and a visuospatial stream of stimuli. Previous studies

comparing the effects of single and dual n-back training have shown that both task versions

seem to be equally effective in producing generalization, and that the transfer effects are fairly

similar (Jaeggi et al., 2010; Jaeggi, Buschkuehl, Shah, & Jonides, 2014). The n-back task is

often highlighted as a WM updating measure (e.g., Szmalec, Verbruggen, Vandierendonck, &

Kemps, 2011), but it also reflects active maintenance and capacity of the WM (plus

interference control if so-called lure stimuli are included in the stimulus sequence; see NIMH

Research Domain Criteria Workshop on Working Memory, 2010). The concurrent validity of

the n-back task has, however, been questioned as previous studies have reported low

correlations between the n-back task and other WM tasks, especially complex span tasks

(Miller, Price, Okun, Montijo, & Bowers, 2009; see Redick & Lindsey, 2013, for a meta-

analysis). Nevertheless, a recent study showed that at a latent level, the n-back task is highly

correlated with other WM tasks (complex spans, WM updating tasks, and sorting spans; Schmiedek, Lövdén, & Lindenberger, 2014).

The efficacy of n-back training has previously been investigated in three recent meta-analyses (Au et al., 2015; Melby-Lervåg & Hulme, 2016; Melby-Lervåg et al., 2016). Au et al. (2015) focused solely on the training effects on Gf. They included twenty studies with healthy adults in their analyses and found a small transfer effect of n-back training to Gf ($g = 0.24$). Au et al. (2015) also investigated the effects of various moderators, such as the type of control group (active or passive), type of n-back task (single or dual), and type of material in the Gf transfer tasks (matrix or nonmatrix; verbal or visuospatial). Active control groups typically receive the same amount of training as the training groups do, but with tasks that are not intended to tap WM to any greater extent. Passive control groups only participate in pre- and postsessions. The results from the Au et al. (2015) meta-analysis indicated that studies with passive control groups showed more transfer to Gf than studies with active control groups. However, their follow-up analyses revealed that this finding did not stem from a difference between active and passive control groups, but from the training groups for some reason performing better in studies with passive controls than in studies with active controls. According to Au et al. (2015), these results do not support the idea that the Hawthorne or expectancy effects affect the results. Finally, their results showed no effects of the other aforementioned moderators. The Au et al. (2015) meta-analysis was challenged by Melby-Lervåg and Hulme (2016) who criticized the exclusion of relevant studies, the calculation of effect sizes without taking pretest differences between groups into account, and the interpretation of the results comparing active and passive control groups. Melby-Lervåg and Hulme (2016) argued that even though the difference between active and passive control groups was not statistically significant in the Au et al. (2015) meta-analysis, there was a difference in the pre-post effect sizes between those two groups. They also emphasized that

the analysis only consisted of 12 studies in each category. Melby-Lervåg and Hulme (2016)

reanalyzed the Au et al. (2015) data by including only the nonverbal reasoning tasks and

correcting for possible pretest differences between groups. They found a very small but

statistically significant effect of transfer to these tasks ($g = 0.13$). Au et al. (2016) have later

responded to this critique, and they maintain that the discrepancy between the effect sizes in

the Au et al. (2015) and Melby-Lervåg and Hulme (2016) analyses is mainly related to

differences in the various meta-analytic decisions, such as the fact that Melby-Lervåg and

Hulme (2016) compared the same treatment group to several control groups without taking

the dependency between these comparisons into account. Furthermore, to avoid confounds

stemming from methodological differences between studies, Au et al. (2016) compared active

control groups to passive control groups within such n-back training studies that had

employed both. The results from those analyses showed no difference between the two types

of control groups.

In their recent meta-analysis on different kinds of WM training, Melby-Lervåg et al.

(2016) included the type of training paradigm as a moderator in their analyses. These results

showed null to moderate training effects from n-back training to untrained WM tasks (verbal

WM: $g = 0.17$ for studies with active controls and $g = 0.12$ for passive controls; visuospatial

WM: $g = 0.24$ with active controls and $g = 0.52$ with passive controls) and very small but

statistically significant transfer effects to nonverbal problem solving ($g = 0.15$ with active

controls and $g = 0.26$ with passive controls). Melby-Lervåg et al. (2016) further pointed out

that those n-back training studies with active control groups that showed the largest transfer

effects to nonverbal ability all had small sample sizes (less than 20 participants per group) and

most of them also showed decreases in performance from pretest to posttest in the control

groups. Melby-Lervåg et al. (2016) emphasized that only studies with active control groups

can be used to evaluate the effectiveness of an intervention and they "recommend that

investigators stop conducting working memory training studies with untreated control groups and that journals stop publishing them" (p. 524).

Compared to the earlier meta-analyses on transfer following n-back training, the present one has several distinctive features that we deem as important. First, as mentioned above, limiting the sources of variability (here training method, age, and health status) should yield results that are easier to interpret. Second, we strictly separated the trained tasks from the untrained ones as mixing these would risk inflating the near transfer effect (see above for the discussion on the meta-analysis by Melby-Lervåg & Hulme, 2013). Third, meta-analytic near transfer measures have lumped together untrained WM tasks that differ from the training tasks only by their stimuli and WM tasks that are structurally different from the training tasks (Melby-Lervåg & Hulme, 2013; Melby-Lervåg et al., 2016; Schwaighofer et al., 2015; Weicker et al., 2016). We believe that one should keep these two transfer measures apart to obtain a more detailed view on the generalizability of WM training. For example, in n-back training studies a near transfer effect rising only from untrained n-back tasks could speak for acquisition of task-specific strategies rather than improved WM per se. Fourth, all previous meta-analyses have included only one measure (one task or average of several tasks) per domain per original sample in the analyses. The rationale for choosing a specific task to represent a certain domain is not always fully explicated. For example, in the meta-analysis by Melby-Lervåg and Hulme (2013), there seems to be some inconsistency in the selection of tasks so that a specific task is excluded for some studies but not for all. In the present meta-analysis, we were able to include all measures from the original studies by employing a multi-level meta-analytical approach, and thereby having less risk of biasing the estimated training effect.
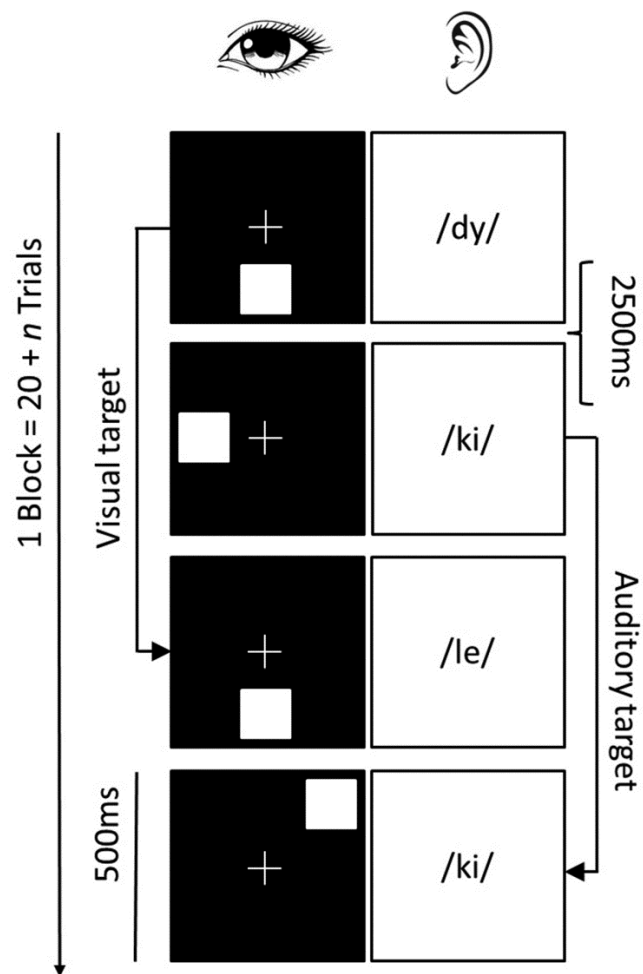
Figure 1. Depiction of the dual n-back task. In a 2-back version, the participants should press the corresponding "yes" button for the third location and for the fourth syllable.

The Structure of the Present Meta-Analysis

In the present meta-analysis on WM training with n-back tasks, we investigated the transfer to untrained tasks representing the following four cognitive domains: task-specific near transfer as measured by untrained versions of the n-back task, WM as represented by other than n-back tasks, cognitive control, and Gf. Cognitive control was included as a transfer domain because earlier studies suggest that WM is closely related to other executive functions (see e.g., Miyake et al., 2000). Transfer to attention and cognitive control has also been studied in two of the previous meta-analyses. Melby-Lervåg and Hulme (2013) included

only the Stroop task in their analyses and Weicker et al. (2016) investigated two executive

domains: cognitive control and executive functioning, and attention and processing speed.

Both of these domains included tasks that can be considered as measures of inhibition (e.g.,

the Stroop task in the former and the Simon and Flanker tasks in the latter one). In our

cognitive control domain, we included tasks that measure inhibition of irrelevant information,

set shifting, multitasking, and verbal executive functions (e.g., verbal fluency). We excluded

tasks that measure attention (e.g., 0-back). Finally, in line with previous meta-analyses we

included fluid intelligence (verbal and nonverbal reasoning) as a separate domain in the

analyses. These four domains also form a rough dimension of the degree of cognitive overlap

between the training tasks and the transfer tasks. If the magnitude of transfer effects follows

the presumed cognitive overlap (Dahlin et al., 2008; Waris et al., 2015), a decreasing degree

of transfer should be seen when moving from n-back transfer tasks to WM, cognitive control,

and Gf. However, as some previous studies suggest that the correlations between n-back tasks

and other WM tasks are very low (e.g., Redick & Lindsey, 2013), it is possible that the

transfer effects to untrained WM tasks are lower in this meta-analysis than in those previous

meta-analyses that investigated various kinds of WM training. This also appears to be the case

in the analyses on n-back training in the Melby-Lervåg et al. (2016) meta-analysis.

Several moderating factors were taken into account in the present meta-analysis,

namely the type of control group (active vs passive), type of training (single vs dual n-back),

training dose (small vs large), transfer material (verbal vs nonverbal), and age (young vs old).

Following the current discussion on the effects of the type of control group (Au et al., 2016),

we compared active and passive controls only within studies (comparing active and passive

control groups within those studies that have employed both). We also used the type of

control group as a co-variate in our analysis of the training effect, thereby considering the

potential influence of the type of control group used in an experiment.

In addition, we implemented some methodological improvements that allowed us to more fully take advantage of the information in the original studies. Since the training of participants is quite resource intensive, a study in this field typically contains a relatively small sample measured on several dependent measures, both before and after training. For a meta-analysis, the information value of a single original study with a small sample is quite limited (for a discussion on this problem in the context of working memory training, see Bogg & Lasecki, 2014), but it can be enhanced by adequate modeling allowing for inclusion of several measures and by utilizing the advantages of the repeated measures design.

Therefore, we made use of a multi-level meta-analytical approach to include effect sizes for all pertinent tasks from the original studies while accounting for the fact that these effect sizes came from the same sample. The multi-level approach has two general advantages. Firstly, it allows us to include more effect sizes and thereby gives us a better picture of the differences in the general training effects between studies. Secondly, it allows for a more complete comparison of improvements on different tasks within the same study, providing data for the aforementioned moderation analyses. Furthermore, we took full advantage of the repeated measures design used in the original studies. A repeated measures design offers higher statistical power than an independent groups design, but only if one accounts for how strongly the repeated measures are correlated. If this is disregarded, as is the case in the previous meta-analyses in this field, the standard errors of the effect sizes are inflated to that of independent groups designs. The effect sizes themselves are not affected, but differences in reliability (for other reasons than sample size) can only be caught by comparing each individual's posttest measure with the corresponding pretest measure. To do this, one needs to analyze the original data or, as in our case, account for the correlation between pre- and posttest measures. Besides more accurately allocating the appropriate relative weight to each study, a more crucial advantage is seen in the increased statistical

power. Thus, by accounting for the correlation between pre- and posttest measures, we raise

the statistical power of our meta-analysis which allows us to make use of the multi-level

approach and more confidently examine moderation effects.

      Our research questions were as follows:

1.   Does n-back training improve performance on other untrained n-back tasks,

          structurally different WM tasks, cognitive control tasks, or fluid intelligence

          tasks?

2.   Is the type of control group (active vs. passive) related to the transfer effects in

          the four domains?

3.   Is the type of n-back training task (single vs. dual) related to the transfer

          effects in the four domains?

4.   Is the effect of dual n-back training different for verbal and nonverbal transfer

          tasks?

5.   Is age related to transfer effects in the four domains?

6.   Is the duration of training related to transfer effects in the four domains?

Table 1

A Summary of the Results from Previous Meta-Analyses Investigating the Effects of Working Memory Training

Study	Training Type	Studies	Sample	WM	STM	Gf	EF	Attention	LTM	Arithmetic	Reading Comp	Word Decoding
Melby-Lervåg & Hulme (2013)	WM	23	Healthy/clinical Children/adults	$d = 0.79$ (V); 0.52 (NV)		$d = 0.19$ (NV); 0.13 (V)	$d = 0.32$			$d = 0.07$		$d = 0.13$
Au et al. (2015)	N-back	20	Healthy adults			$g = 0.24$						
Melby-Lervåg & Hulme (2016)	N-back	17-23	Healthy adults			$g = 0.13$ (NV)						
Weicker et al. (2016)	WM	103	Healthy/clinical Children/adults	$g = 0.44$		$g = 0.24$	$g = 0.20$	$g = 0.20$	$g = 0.18$			
Schwaighofer et al. (2015)	WM	47	Healthy/clinical Children/adults	$g = 0.55$ (V); 0.63 (NV)	$g = 0.37$ (V); 0.72 (NV)	$g = 0.16$ (V); 0.14 (NV)				$g = 0.09$		$g = 0.08$
Melby-Lervåg et al. (2016)	WM	87	Healthy/clinical Children/adults	$g = 0.31$(A; V); 0.42 (P; V)/0.28 (A; NV); 0.51 (P; NV)		$g = 0.05$ (A; NV); 0.20 (P; NV) /0.05 (A; V); 0.03 (P; V)				$g = 0.06$ (A); 0.12 (P)	$g = 0.15$ (A); 0.12 (P)	$g = 0.08$ (A); 0.01 (P)
	N-back	32	Healthy/clinical Children/adults	$g = 0.17$ (A; V); 0.12 (P; V)/ 0.24 (A; NV); 0.52 (P; NV)		$g = 0.15$ (A; NV); 0.26 (P; NV)/0.04 (A; V); 0.02 (P; V)						

Note. WM = working memory; STM = short-term memory; Gf = fluid intelligence; EF = executive functions; LTM = long-term memory; Reading comp = reading comprehension; V = verbal; NV = nonverbal; A = compared to the active control group; P = compared to the passive control group

Method

Literature Search and Criteria for Inclusion

We searched the electronic databases PubMed, PsycINFO (ProQuest), Google

Scholar, Google, and ERIC for relevant research reports (articles, dissertations, and theses).

The first search was conducted in February 2015. The search term *n-back training* was used

in all databases. In Google Scholar and Google, citation marks (*"n-back training"*) were used

in order to reduce noise in the search. The first search resulted in 99 hits in PubMed, 86 in

PsycINFO, 189 in Google Scholar, 12,500 in Google, and five in ERIC. We screened all hits

in the databases with the exception of Google, where we limited screening to include only the

first 100 hits. For a study to be included at this stage, the following criteria had to be met:

1. A randomized controlled trial with at least one training group and at least one
   control group and a pretest-posttest design.

2. Transfer tests were employed within any of the following cognitive domains:
   n-back, WM, fluid intelligence, and cognitive control.

3. The training program included only n-back task(s).

4. The training task in the active control group was not a WM task.

5. The sample consisted of healthy adults.

Search hits were screened in the aforementioned order of the databases. Our inclusion

criteria yielded 12 studies from PubMed, eight additional studies from PsycINFO, 12

additional studies from Google Scholar, and two additional studies from Google (all numbers

refer to unique studies; the number of overlapping studies is not reported). Our search thus

resulted in 34 reports. In the next step, three theses were excluded because the data was also

included in published articles, and one thesis was excluded because different transfer

measures were used before and after training. Two theses had been written on the same data

collection and were therefore treated as one study. The initial literature search thus resulted in 29 studies.

The literature search was updated three times; once in July 2015, once in February 2016, and once in October 2016. The search procedure was the same as in the initial literature search. These searches resulted in the inclusion of six additional studies, one from the first search, one from the second search, and four from the last search. The reference lists of the included studies and prior meta-analyses in the field were also screened, and as a result, six potentially interesting studies were found. Moreover, a yet unpublished study conducted by two of the authors (Soveri, Karlsson, Waris, GrönholmNyman, & Laine, submitted) of the present meta-analysis was included.

In sum, the total number of studies identified for inclusion was 42. The authors of 18 of these studies were contacted in order to receive necessary information not reported in the published articles. We received the requested data for nine of these. However, the remaining nine articles had to be excluded as we did not receive the information, or the authors informed us that the requested data was reported in other articles already included in our sample of studies. Therefore, our final sample consisted of 33 studies. Figure 2 depicts a flow chart of the screening process.

Figure 2. Flowchart depicting the search for literature and the screening process.

We have chosen to exclude some of the n-back training studies with healthy adults that have been included in previous meta-analyses. These studies are listed in Table S1 together with the reason for exclusion. Furthermore, the Seidler et al (2010) study in Au et al (2015) is referred to as Anguera et al (2012) in the present meta-analysis.

Data Coding

Type of group. The type of group was described with a variable with three levels: "Training group", "Active control group", and "Passive control group". The criterion for

considering a control group as an active control group was that participants received the same amount of training as the training group, but that, purportedly, their training program did not load on WM. Passive control group participants did not receive any intervention and participated in the pre- and posttests only.

Training task. The training tasks were described by two variables. "Training type" described whether a single or dual (or, in one case, a triple) n-back task was employed. In a single n-back task, the participants responded only to one stream of stimuli, while the dual n-back task consisted of two streams that were presented simultaneously. "Training materials" defined whether the training task had been verbal, nonverbal or both verbal and nonverbal. The n-back task was considered verbal when the stimuli consisted of language materials such as numbers, letters, or syllables. Task versions without verbal elements were coded as nonverbal. Those training tasks that included both verbal and nonverbal stimuli were exclusively dual n-back tasks.

Training dose. Training dose was operationalized with two variables: "Training hours" and "Training sessions". The former one corresponded to the total number of hours the participants had been training and the latter one to the total number of training sessions. Due to an uneven distribution, we dichotomized the two variables based on a median-split. For the amount of training, the levels were: 6.67 hours or less and more than 6.67 hours. When the information about training dose was imprecise (e.g., 20-25 minutes per session), the lowest number was used to calculate the total amount of training. The median-split for training sessions was as follows: 15 sessions or less and more than 15 sessions.

Transfer tasks. Transfer was described by two variables: "Transfer domain" and "Transfer materials". "Transfer domain" assigned transfer tasks to one of the following cognitive domains: n-back, WM, cognitive control, and Gf. All untrained versions of the n-back task were coded as n-back transfer measures, that is, measures of task-specific transfer.

The remaining WM tasks employing some other paradigm than n-back were coded as WM transfer tasks. Here we included for example simple span tasks (e.g., digit span), complex span tasks (e.g., operations span), and running memory span tasks (e.g., letter-memory). In our cognitive control domain, we included tasks that tap inhibition of irrelevant information (e.g., Stroop and Simon task), set shifting (e.g., Number-letter task), multitasking (e.g., SynWin), and verbal executive functions (e.g., verbal fluency). Tasks that can be considered as measures of attention and vigilance were not included in the cognitive control domain. Finally, all tasks where the participant was required to solve problems without being able to rely on previous experience to any greater extent, were coded as measures of Gf. Reading comprehension was not included in the Gf measure due to the fact that in those tasks, the correct answer is available in the testing materials. The few transfer tasks that did not fit into these categories were excluded from the analyses due to the low number of tasks per remaining domain. One of these was the 0-back condition of the n-back task which was not included as an n-back measure because it exerts only a minimal load on WM. Neither was it coded as a measure of cognitive control, as it is a vigilance task and a variant of the trained task.

For some tasks, several measures were reported. For those tasks for which both reaction times (RT) and accuracy were available, we decided to code only the RT measures as on some tasks accuracy rates are plagued by ceiling effects. Also, for such tasks that included several conditions (e.g., single task trials, repetition trials, and switching trials in set shifting tasks), we decided to include difference scores (e.g., the switching cost, which is the difference between switching trials and repetition trials) when possible. In case these were not available, we included the conditions that involved executive load (i.e., the switching condition).

In order to investigate whether transfer was more likely to the verbal or nonverbal domain, the variable "Transfer materials" was created. Here, the two transfer domains "Gf" and "WM", for which we obtained the largest number of comparisons, were further divided into verbal vs. nonverbal reasoning and verbal vs. nonverbal WM.

The coding of transfer tasks into cognitive domains (i.e., task-specific transfer as measured by n-back, WM, cognitive control, and Gf), and materials (i.e., verbal vs. nonverbal), was based on consensus decisions by three of the authors (A.S., M.L., and L.K.).

Age. Age was dichotomized due to data-availability issues (mean ages were reported in some studies and age ranges in others) and because of its non-normal distribution. The variable "Age group" was coded in the following manner: 59 years or younger and 60 years or older. This division was made based on the age ranges reported in the studies. For studies that did not include information about age range, the decision was made based on other kinds of information provided in the studies (e.g., university students and/or the mean age and its standard deviation for the sample).

Furthermore, we coded the publication year of the articles and theses and the countries where the studies were conducted. More detailed information about the abovementioned variables in each original study is given in Tables S2 and S3.

Statistics for effect size calculation. Group means, standard deviations, and group sizes were extracted for each group to calculate a measure of gain score for each group per transfer measure. To obtain the correlations between pre- and posttest performance (such correlations were reported for all measures in only one of the studies), we contacted the authors of the original studies. After the authors had responded, we had correlations for 406 of the 507 within-sample pre-post comparisons. Where authors did not provide the correlations, we attempted to deduce the correlation from other available information in the study (e.g. the means and standard deviations in combination with a *t* or *F*-value). We could deduce or give a

conservative estimate for 80 out of the 101 remaining pre-post correlations. Where we were

unable to estimate the correlation, we entered 0 as a correlation, essentially treating the pre-

and posttest measures as independent groups. The effect of using a conservative correlation is

that the within-sample pre-post comparison is given a larger standard error and therefore less

weight in the meta-analysis. It is appropriate that we give less weight to studies where we are

less certain about the information provided in the summation of effect sizes, but overly

conservative estimates for the standard errors cannot be included in analyses where the

standard errors are of primary interest, such as a publication bias analysis. A meta-analysis of

the pre-post correlations showed that the correlations were $r = .659$ ($k = 203$, p $< .001$) for the

pre-post measures in the training group, $r = .649$ ($k = 94$, p $< .001$) for the pre-post measures

in the active control group, and $r = .656$ ($k = 109$, p $< .001$) for the pre-post measures in the

passive control groups.

To address the reliability of our coding of group sizes, means, and standard deviations,

we conducted a test of inter-rater reliability. Ten of the 33 studies were randomly selected and

re-coded by an independent researcher and the percentage of agreement (i.e., same value or

not) was calculated. The interrater agreement was 97.1% for the 34 group size values, 100.0%

for the 284 mean vales, and 99.6% for the 284 standard deviation values in the ten studies.

Closer scrutiny revealed two points of disagreement. When we returned to the studies to

retrieve the correct information, we found that the first disagreement was an error made by the

independent researcher. Because the second disagreement amounted to less than one decimal

point, we deemed the coding as reliable.

Meta-Analytical Procedures

All effect sizes were calculated and all meta-analyses were conducted using the

Metafor package (Viechtbauer, 2010) for R (R Core Team, 2008).

Calculation of effect sizes. The effect sizes of interest were pairwise comparisons of pre-post changes (i.e., the difference in performance between the first and second cognitive test sessions) between a training group and a control group. To obtain these effect sizes, we first retrieved the within-group pre-post effect sizes (g_{WITHIN}) for each group and measure included in a study. For g_{WITHIN}, we calculated the standardized mean pre-post change for each training and control group separately using the *escalc* function in the *metafor* package (Viechtbauer, 2010). In the function documentation, this effect size is called a "standardized mean change using raw score standardization" (SMCR). Here, the raw gain score on a task is standardized by dividing it by the standard deviation of the pretest score, and the variance of the effect size is adjusted by accounting for the correlation between the pre- and posttest measure. To make the g_{WITHIN} effect sizes comparable between training and control groups, we first pooled the standard deviation across all groups within the same experiment.

To calculate an effect size representing *the training effect* ($g_{BETWEEN}$), that is, a between-groups effect size representing the difference in pre-post change between a training group and a control group, we subtracted the g_{WITHIN} for the control group from the g_{WITHIN} for the training group. A positive value thus indicates that the pre-post improvement was larger (or, that the pretest-posttest decline was smaller) in the training group compared to the control group. The variance for ($g_{BETWEEN}$) was calculated by summing the variance of the within-group effect size for the two groups.

Multi-level modeling. Six of the 33 studies (i.e., articles, theses, or manuscripts) contained more than one experiment. Here, separate experiments have their own set of training and control groups. For example, one experiment might compare a training group of young individuals to a control group of the same age, while another experiment compares training and control groups with older individuals. Because one of the studies contained four experiments and five studies contained two experiments, a total of 41 experiments were

included from the 33 studies. Experiments also varied in the number of training and/or control groups included. Twelve experiments (in 11 studies) included more than one training group. Eight other experiments (in six studies) involved both an active and a passive control group. The 41 experiments included, in total, 103 different training or control groups. These groups had been measured before and after training, on between one and 13 different measures, depending on the study and experiment. In total, the meta-analysis contained 507 within-sample pre-post comparisons (g_{WITHIN}) from a total of 2105 participants.

These 507 within-group effect sizes yielded 305 between-groups effect sizes (i.e., comparisons between a training group and a control group within an experiment). There were, however, several forms of dependency between effect sizes in this structure. Effect sizes that are dependent are generally less informative than independent ones. When two measurements are perfectly correlated (e.g., measuring the improvement in WM with two or more entirely interchangeable measures), these measurements will provide the same information. This means that including interchangeable measurements, while assuming that they provide independent and unique information, unduly inflates the confidence in the results of a meta-analysis. In other words, unwarranted repetition of the same information will lead to an incorrect narrowing of the confidence estimates in a meta-analysis, which, in turn, increases the risk of making type I errors (Becker, 2000; van den Noortgate, López-López, Marín-Martínez, & Sánchez-Meca, 2013).

The following forms of dependency between effect sizes need to be considered here: The most obvious dependency is that when multiple tasks are used to measure the training effect within the same experiment, the outcomes on these tasks are likely correlated. In other words, these outcomes represent the training effect within the same participant(s). Second, effect sizes for a training effect are also correlated when a particular training group is compared with multiple control groups within an experiment. Third, and similarly, effect sizes

are correlated when multiple training groups are compared with one control group. Fourth, and lastly, effect sizes from multiple experiments within a particular study may be correlated due to similarities in experimental methodology and the experimental setting.

With an adequate amount of data, the correlation within each of these levels of dependency can be estimated in a multi-level framework. This model is, however, too complex for the amount of data available, but a few simplifications can be made that will make modelling considerably more parsimonious with little loss of information and with few added assumptions: (a) The experiments in the four studies that had more than one experiment were too few to allow us to estimate the variance between experiments within studies. All experiments, however, have separate participants and the dependency within participants is therefore handled adequately. Thus, for the sake of parsimony we did not model the dependency between experiments within the same study, but treated all 41 experiments as independent samples; (b) When there were more than one training group, we performed an internal fixed effects meta-analysis of the g_{WITHIN} in the groups producing a single g_{WITHIN} for the training groups in that experiment, and in the experiments where both an active and a passive control group was used we disregarded the passive group (except for the comparisons between control groups within studies). These simplifications limit each experiment to one single comparison between a training group and a control group per measure. This eliminates dependency between comparisons within an experiment that would reuse training or control groups.

What remained were 203 estimated training effects (i.e., comparisons of a training vs. control group) in 41 separate experiments. These comparisons, in this framework, are presented in Figure 3 (see Table S3 for more detailed information about pre-post effect sizes and group comparisons). Hence, our model is a three-level model with random intercepts for each unique experiment and for each unique measure within those experiments. In other

words, in addition to observing the variance of participants' results on a given measure in a

given experiment, we estimate the variance of the mean results of different experiments, and

the variance of mean results for different measures within an experiment. Thereby, we can

account for, and estimate, the dependency of effect sizes between-measures within

experiments.
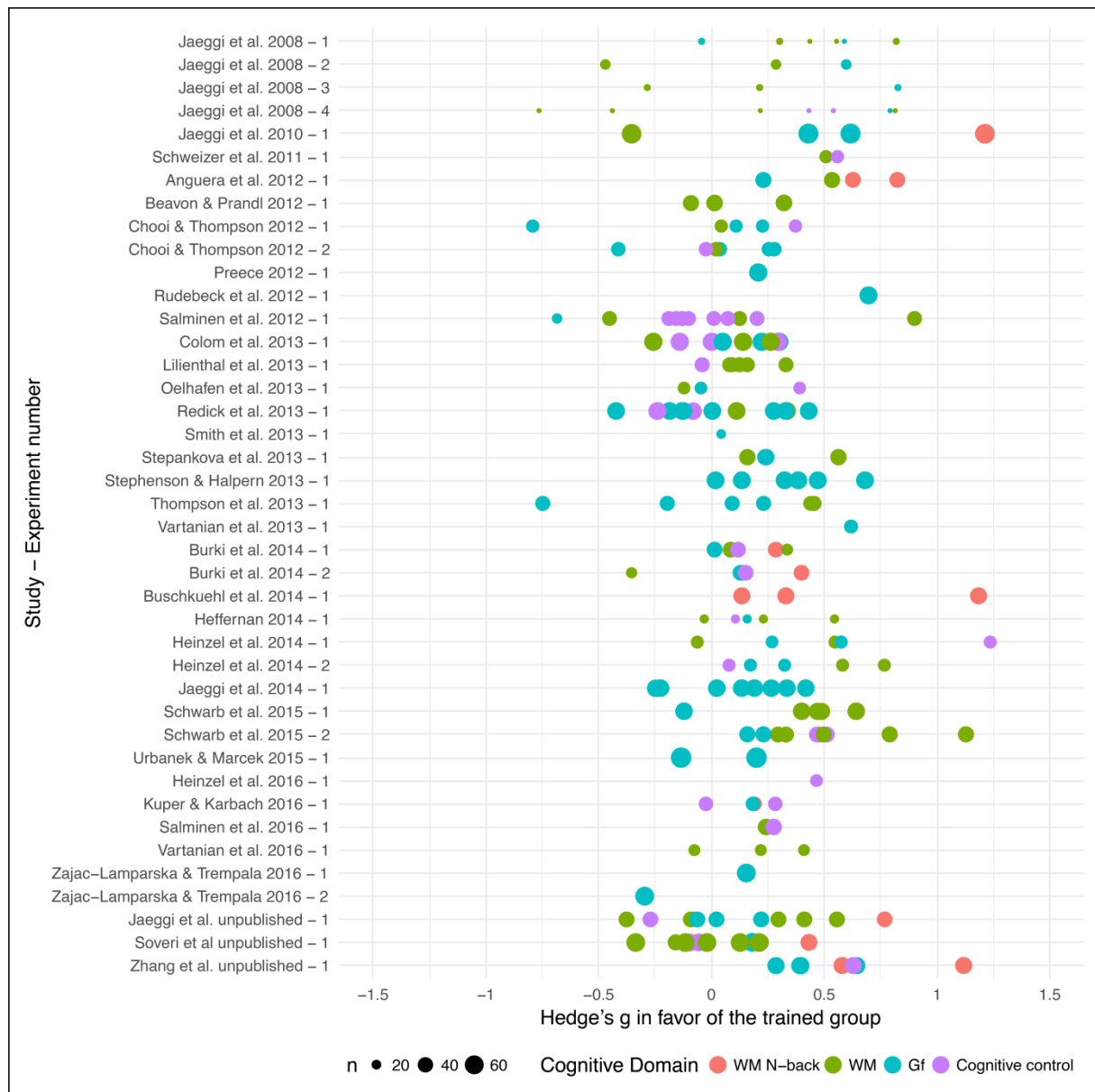
Figure 3. Schematic overview of the multi-variate and multiple-treatment structure of the

included studies. The figure displays effect sizes, sample sizes, and cognitive domain

(untrained n-back tasks [WM N-back], Working Memory [WM], Cognitive control, and Fluid

Intelligence [Gf]). The horizontal axis indicates a Hedge's *g*, with positive effects (favoring training groups over control groups) to the right and vice versa. Two effect sizes are outside the plotted range.

This dependency, or how strongly measurements are clustered within experiments, can be estimated through the *intraclass correlation coefficient* (*ICC*). It can be estimated by dividing the variance between experiments by the sum of the variance between and within experiments. The *ICC*, thus, also describes the amount of the variance in the effect sizes that can be attributed to differences between experiments. In Figure 3, variance both between and within experiments can be discerned. When the variance within studies is small in relation to the variance between studies, the correlation within studies is deduced to be high. In contrast, if measures within studies vary to the point where each measure could just as well belong to any one of the experiments, the correlation is deduced to be close to zero. The *ICC* for the clustering of training effect among measurements within an experiment was .499. To evaluate whether the added model complexity paid off, we also conducted ANOVA tests. The test between the single- and two-level models was significant (*LRT* = 24.34, $p < .001$), and so was the test between the two- and three level models (*LRT* = 12.34, $p < .001$). This indicates that the three-level model is more adequate than a model treating all measurements as independent.

Publication bias. Several methods have been suggested for the challenging task of identifying and estimating the magnitude of publication bias (Rothstein, Sutton, & Borenstein, 2005). One of the most often used among these methods is the trim and fill method (Duval & Tweedie, 2000). However, more novel and promising methods, such as *p-curve* (Simonsohn, Nelson, & Simmons, 2014) and variations of Egger's regression (Egger, Smith, Schneider, &

Minder, 1997; Moreno et al., 2009) have been found to perform better than the trim and fill method.

The application of all these methods is complicated by the dependency between effect sizes and the multi-level approach in this study. For example, common p-curve methods require independent effect sizes (Simonsohn et al., 2014). Although a method accounting for dependency has been recently suggested (Bishop & Thompson, 2016), p-curve methods are problematic if studies include ghost variables (i.e., outcome measures that have not been reported in the final study). While selecting a subset of independent effect sizes could relieve the problem of dependency, this approach would lead to loss of information without resolving the problem of ghost variables. For our purposes, we find a careful visual inspection of funnel plots accompanied by regression-based methods to assess funnel plot asymmetry to be the most fruitful approach.

We find two forms of funnel plots especially informative for our purposes. One is the so-called *contour enhanced funnel plot* (Peters, Sutton, Jones, Abrams, & Rushton, 2008) where each effect size is plotted against the precision (inverse of the standard error). The vertical reference line is drawn at 0 and the contours change shade at different levels of two-tailed *p*-values. When publication bias is absent, the individual effect size measures should be distributed symmetrically around the mean effect size, with measures with higher precision distributed more closely around the mean effect size. Visually, this typically forms an inverted funnel with the tip of the funnel further up in the plot and close to the mean effect size. Publication bias due to underreporting of non-significant findings should result in an asymmetrical funnel with effect size estimates missing in the area of the plot indicating nonsignificant results.

The second form of funnel plot that we find informative we call a funnel plot of residuals. Publication bias is only one possible explanation for funnel plot asymmetry.

Another feasible explanation is systematic heterogeneity due to moderators (Egger et al.

1997). By plotting the residuals for the effect sizes in the model against the standard error

(with lower standard error higher on the y-axis), the moderators in the model can be

accounted for and the funnel plot asymmetry in relation to the expected value for each effect

size measure can be examined. Similar contours as in the contour enhanced funnel plots can

be added. In these plots, we included the between-study and the between-measure variance in

the total variance for an effect size. We used this plot to compare the effect sizes to the mean

effect for each cognitive domain. Simultaneously, we accounted for the variance between and

within studies. Thereby, we were able to use these plots to also examine outliers.

To more formally test for funnel plot asymmetry, we added the standard error or

variance of the effect size as a predictor to pertinent meta-regression models. This is a close

equivalent of the PET-PEESE method suggested by Stanley and Doucouliagos (2014), while

simultaneously retaining our three-level model.

In the PET-PEESE approach the effect size is first regressed on the standard error of

the effect size in a weighted least squares (WLS) regression with the standard error of the

effect size as the weight. Testing if the slope of the regression line is statistically significant

serves as a test for statistically significant publication bias. The typical relationship when

publication bias is present is that higher standard errors are associated with larger effect sizes.

The intercept of this regression is interpreted as an estimate of the effect in a hypothetical

study of no error (SE = 0) and therefore no bias. Testing whether the intercept is statistically

significant serves as a test of whether there is a true, bias corrected, effect different from 0.

This test is called the *precision-effect test* (PET). In the simulation study by Stanley and

Doucouliagos (2014), PET performed well when the true effect in the meta-analyzed studies

was 0, with the intercept being a slight overestimation of the true effect. When there was a

true effect, however, using the variance as the predictor in the regression showed better

performance. Here the intercept was a slight underestimation of the true effect. This test is

called a *precision-effect test with standard error* (PEESE). The authors thus suggested that

PET is followed up by PEESE if PET shows a true bias corrected effect.

Stanley and Doucouliagos (2014) did not evaluate the method in a three-level model

but we find this approach to be the best available method to estimate the true effect adjusted

for possible publication bias. Furthermore, we find that adjusting for pertinent moderators in

the same model is motivated to exclude the possibility that there are small-study effects

associated with the moderators rather than explainable by publication bias. As such, this is a

test most closely related to the funnel plot of the residuals.

We did not plot effect sizes where we could not estimate the pre-post correlation, as

our conservative estimate (0) unduly inflates the standard error and those observations thus

differ from the other combinations of effect sizes and their standard errors.

Results

Descriptive Results

The study sample consisted of 33 studies, including 41 experiments, with, in total, 103

experiment groups. Of these groups, 54 (52.4%) were training groups and 49 (47.6%) were

control groups. Twenty-three (42.6%) of the training groups practiced with a single n-back

task, 30 (55.6%) with a dual n-back task, and one (1.9%) with a triple n-back task. Twenty

(40.8%) of the control groups were active and 29 (59.2%) were passive. See Table 2 for a

summary of study characteristics.

Table 2

Number of Experiments and Number of Training-Control Comparisons per Study Characteristic

Study characteristic	Number of experiments	Number of comparisons
Total	41	305
Training groups		
Single n-back	20	115
Dual n-back	25	188
Triple n-back	1	2
Control groups		
Passive	29	198
Active	20	107
Cognitive domain		
N-back	9	17
Working memory[a]	28	104
Verbal	28	70
Nonverbal	13	29
Fluid intelligence	35	133
Verbal	4	15
Nonverbal	35	118
Cognitive control	23	51
Sample age[b]		
Young	33	271
Old	5	27
Year		
2008	4	20
2010	1	8
2011	1	6
2012	7	44
2013	9	94
2014	7	54
2015	3	29
2016	6	21
Unpublished	3	29
Country of origin[c]		
Australia	2	4
Canada	3	9
Czech Republic	2	12
Finland	1	11
Germany	6	33
Poland	2	6
Spain	1	9
Switzerland	8	47
Taiwan	1	8
UK	2	8
USA	13	158

Note. Number of comparisons represents training-control comparisons for each measure. Young = sample age under 60; Old = sample age 60 or over.
[a]For five of the comparisons in two experiments, the working memory task could not be categorized as verbal or nonverbal. [b]Six samples in three experiments (seven comparisons) could not be categorized as either young or old. [c]The unpublished manuscript by the first author of the present meta-analysis is excluded due to masked review.

Data Screening and Assessment of Publication Bias

Firstly, we investigated publication bias separately for each cognitive domain. We plotted the effect sizes in four separate contour enhanced funnel plots – one for each domain. We then fitted a model with cognitive domain as a moderator and plotted the residuals in four separate funnel plots of residuals.

In the contour enhanced funnel plots (Figure 4, left column), each effect size is plotted against its precision (1/SE, the inverse of the standard error). The vertical reference line is at g = 0. Contours change shade at different levels of two-tailed p-values. Publication bias due to underreporting of non-significant findings should result in an asymmetrical funnel with effect size estimates missing in the area of the plot indicating nonsignificant results. While there was some asymmetry, there was no obvious pattern of statistically significant effect sizes being preferred over nonsignificant effect sizes.

In the funnel plot of residuals (Figure 4, right column), each effect size is plotted against the SE with lower SE higher up. The vertical reference line is at the aggregated effect size for the pertinent cognitive domain. The effect sizes are thus plotted in relation to the expected value on each domain respectively, with outliers falling outside the contour lines. Cognitive domain was indeed a statistically significant moderator, Q_M (3) = 21.60, $p < .001$.
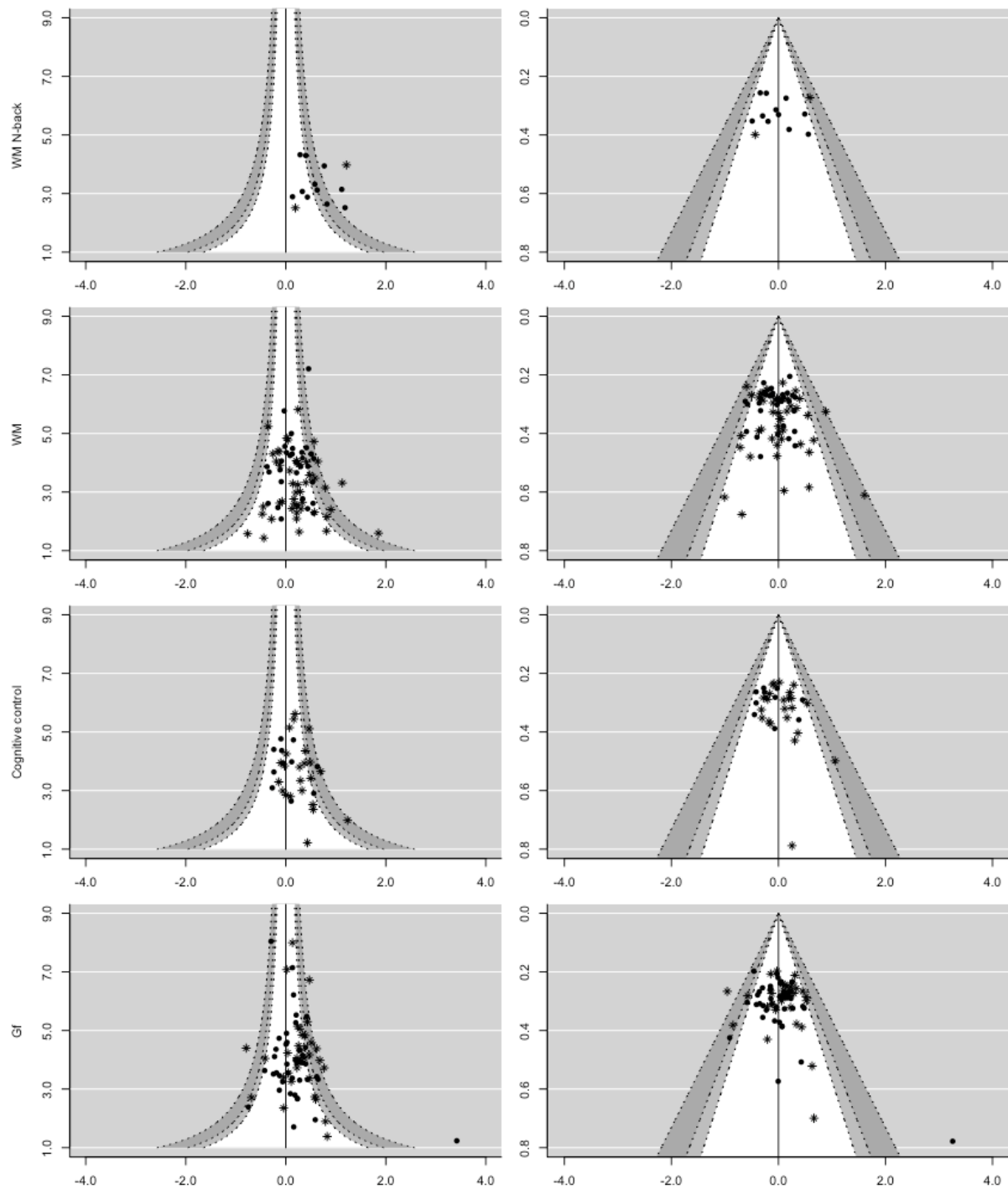
Figure 4. Contour enhanced funnel plots for each cognitive domain (by rows). In the funnel plots in the left column, effect sizes are plotted against their precision (1/*SE*). In the funnel plots in the right column, effect sizes are plotted against the *SE*. The reference lines represent an effect size of Hedge's $g = 0$ (left column) and the mean Hedge's g within each domain (right column).

In these funnel plots asymmetry is a little more apparent. For cognitive control and Gf there is a group of effect sizes with high *SE* that show higher effect sizes than the aggregated effect size for that domain. Furthermore, for all domains taken together there is a group of effect sizes that stands out because their *SE* is uncharacteristically high for this sample. An examination revealed 12 effect sizes with standard errors higher than 0.5, which proved a natural cut-off point. Note that the standard error in the funnel plot is based on the combined variance from each level of the multilevel model and thus does not completely correspond to the standard error calculated for the effect size alone.

There are two reasons for excluding these 12 effect sizes from further analyses. First, removing them markedly reduces funnel plot asymmetry. Effect sizes with high standard errors are indeed most susceptible to publication bias and removing them will only lead to a small decrease in statistical power. Second, these effect sizes are outliers in respect to their standard error and thereby they risk having a strong influence on PET-PEESE analyses, in which the standard error is used as a predictor.

The funnel plot of residuals also reveals two effect sizes for Gf that we considered outliers. These two had standardized residuals of 3.9 and -3.4 respectively. One of these was also excluded based on its large standard error. We thus excluded 13 out of the originally 203 effect sizes (See Table S3 for these exclusions).

Training Effects by Cognitive Domain

We then investigated the transfer effects by cognitive domain. We found that the effect was largest for n-back, $g = 0.62$, [0.44, 0.81], $p < .001$, followed by WM, $g = 0.24$, [0.16, 0.32], $p < .001$, cognitive control, $g = 0.16$, [0.05, 0.27], $p < .001$ and Gf, $g = 0.16$, [0.08, 0.24], $p < .001$ (See Figure 5).

Cognitive Domain	Control Groups	Training Effect	Hedge's g [95% CI]
WM N-back	11 Active, 2 Passive		0.62 [0.44 , 0.81]
WM Other	31 Active, 42 Passive		0.24 [0.16 , 0.32]
Cognitive Control	11 Active, 22 Passive		0.16 [0.05 , 0.27]
Gf	38 Active, 32 Passive		0.16 [0.08 , 0.24]

-0.20   -0.00   0.20   0.40   0.60   0.80   1.00

Favors Control Group   Hedge's g   Favors Training Group
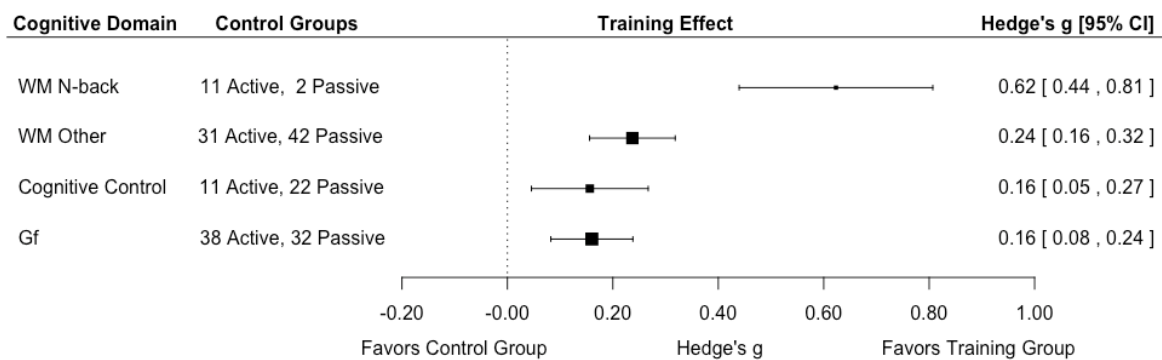
Figure 5. For each cognitive domain, the figure displays synthesized effect sizes and confidence intervals for the comparison of pretest-posttest improvements between training groups and control groups. Positive values favor the training groups and negative values favor control groups.

The moderation effect was significant Q_M (3) = 23.77, $p < .001$. The effect on n-back tasks differed from effects on all other tasks (all ps < .001). There were no significant differences between the effects for other domains (all ps ≥ .134). The test for residual heterogeneity was also significant Q_E (186) = 230.59, $p < .05$.

After this we reran our main analysis with all data included. Again, the effect was largest for n-back, $g = 0.63$, [0.44, 0.82], $p < .001$, followed by WM, $g = 0.24$ [0.16, 0.33], $p < .001$, cognitive control, $g = 0.17$, [0.06, 0.29], $p < .001$, and Gf, $g = 0.16$, [0.08, 0.24], $p < .001$. Note that although we find the exclusion of outliers and effect sizes with uncharacteristically high effect sizes beneficial to the precision of our analyses, the results are very similar before and after exclusion.

As a test for potential small-study effects that could remain even after our data screening, we added first the standard error and then the variance of the effect size as a

predictor in PET-PEESE type analyses. Neither the standard error ($b = 0.067$, $SE = 0.338$, $p =$ .844) nor the variance ($b = -0.037$, $SE = 0.599$, $p = .951$) were statistically significant predictors of effect size. Because the estimated effect sizes for the cognitive domains were close to identical to the ones reported above in both analyses (the largest discrepancy was 0.02 points), we conclude that, after data screening, there was no evidence of publication bias affecting our conclusions.

Control Analyses

The type of control group (Active vs. Passive), moderated slightly, but statistically significantly, the estimated training effect ($g = 0.11$, [0.00, 0.22], $p = .048$). After controlling for the type of control group, training effects were still largest for n-back, $g = 0.59$, $p < .001$, followed by WM, $g = 0.18$, $p < .01$. The effect sizes for cognitive control was $g = 0.08$, $p =$ .202. For Gf the effect size was $g = 0.11$, $p > .05$. (Estimates for experiments with passive control groups are 0.11 points higher).

The analysis on transfer to WM was run also without simple spans so that the WM domain would be more similar to previous meta-analyses that have investigated simple spans separately. This resulted in a slightly smaller training effect for WM ($g = 0.21$, [0.12, 0.30], p < .001), but the effect size was within the confidence intervals of the earlier estimate.

We also reran the analysis with n-back measures pooled with other measures of WM (including simple spans). This resulted in an increase of the training effect on WM, $g = 0.29$ [0.21, 0.37], $p < .001$.

Efficacy of N-back Training in Different Training Contexts

Furthermore, we investigated the effects of training type (single or dual n-back training), training dose (number of sessions and hours of training), and sample age (young or old) on the effects of N-back training. Training type (single or dual n-back training) did not produce significantly different training effects, $Q_M (1) = 2.19$, $g = 0.09$ [-0.03, 0.21], $p = .139$,

as there was a similar training effect for single n-back training, $g = 0.27$ [0.18, 0.37], $p < .001$, and dual n-back training, $g = 0.18$ [0.11, 0.26], $p < .001$. The test for residual heterogeneity remained significant, $Q_E (188) = 254.86$, $p < .001$. Because the main analyses showed a significant effect of cognitive domain, we reran the analysis with cognitive domain as a covariate. This did not alter the effect of training type, $g = 0.06$ [-0.05, 0.18], $p = .300$.

We did not find any effect of the number of sessions, $Q_M (1) = 0.02$, $g = -0.01$ [-0.13, 0.12], $p = .902$, and $g = -0.03$ [-0.14, 0.09], $p = .685$, with cognitive domain as a covariate. Neither did we find any effect of hours of training, $Q_M (1) = 0.03$, $g = 0.01$ [-0.12, 0.14], $p = .872$, and $g = g = 0.00$ [-0.12, 0.12], $p = .981$ with cognitive domain as a covariate.

We then analyzed whether sample age moderated the training effect. We found no effect of age, $Q_M (1) = 0.07$, $g = 0.03$ [-0.17, 0.22], $p = .794$. Adding cognitive domain as a covariate to the analysis did not alter the effect, $g = 0.02$ [-0.16, 0.20], $p = .834$.

Transfer of N-back Training to Different Materials

To investigate whether dual n-back training produced different effects depending on the transfer task materials (verbal vs. non-verbal), we conducted an analysis in which only WM and Gf measures were included. The n-back measures and cognitive control measures were not included because the number of effect sizes was insufficient. The difference between verbal and non-verbal transfer measures was not significant, $Q_M (3) = 1.77$, $p = .623$. The test for residual heterogeneity was not significant, $Q_E (103) = 123.43$, $p = .0.08$ (See Figure 6).

Cognitive Domain	Control Groups	Training Effect	Hedge's g [95% CI]
WM Non-verbal	8 Active, 13 Passive		0.19 [0.04 , 0.34]
WM Verbal	22 Active, 28 Passive		0.18 [0.08 , 0.28]
Gf Non-verbal	31 Active, 31 Passive		0.18 [0.10 , 0.26]
Gf Verbal	7 Active, 1 Passive		0.03 [-0.19 , 0.25]

-0.20   -0.00   0.20   0.40   0.60   0.80   1.00

Favors Control Group   Hedge's g   Favors Training Group
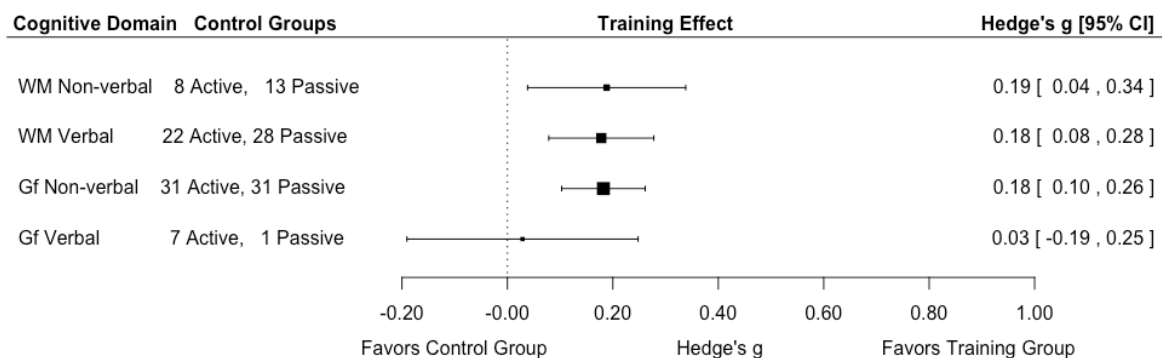
Figure 6. For both verbal and nonverbal transfer task materials for fluid intelligence (here: reasoning) and working memory (WM), the figure displays synthesized effect sizes and confidence intervals for the comparison of pre-post improvements between training groups and control groups. Positive values favor the training groups and negative values favor control groups.

Active and Passive Control Groups

To investigate whether the pre-post change in active and passive control groups differed, we conducted a comparison between active and passive control groups within the same experiment. Here g is calculated similarly to the treatment effect; the difference in g_{WITHIN} between the active and the passive control group. This analysis of, in total, 34 effect sizes revealed no differences between the two types of control groups, $g = 0.03$, [-0.07, 0.12], $p = .429$. The effect of the type of control group was not significantly different between the four cognitive domains, $Q_M (3) = 6.81$, $p = .078$.

Discussion

We set out to examine the efficacy of one of the most commonly used WM training paradigms, namely the n-back task. In contrast to previous meta-analyses in the field, we did not lump untrained versions of the training task together with the other WM transfer

measures. Instead, we deemed it as important to keep these two types of transfer measures separate to be able to examine whether n-back training mainly yields task-specific transfer or more general WM improvement. We also took into account certain methodological shortcomings that limit the interpretability of previous meta-analyses on WM training effects. From each original study sample, previous meta-analyses have included only one measure (or an average of several measures) per cognitive domain in the analyses. We employed a multi-level meta-analytical approach to be able to include all measures from the original studies, and thereby obtain a less biased estimate of the training effects. We also took full advantage of the repeated measures design in the original studies by accounting for the correlation between pre- and posttest performance and thereby increasing the statistical power of our analyses.

Main Analyses

The present meta-analysis included 203 training effects (190 after data screening) from 33 studies. These studies consisted of 41 unique experiments. In total, data was obtained from 2105 individuals. The results from the main analyses showed a moderate effect of task-specific transfer to untrained n-back tasks and very small transfer effects to other untrained WM measures, cognitive control, and Gf.

The transfer effect to untrained WM tasks as a whole is approximately of the same size as in the n-back training analyses by Melby-Lervåg et al. (2016; when averaged across verbal and nonverbal WM and active and passive control groups), but considerably smaller than in the other meta-analyses not focusing specifically on n-back training (Melby-Lervåg & Hulme, 2013; Weicker et al., 2016; Schwaighofer et al., 2015; Table 1). There are several possible reasons for this discrepancy. First, Melby-Lervåg & Hulme (2013) who reported the highest effects of near transfer, in some cases included the training task in their analyses, resulting in an overestimation of the near transfer effect. Second, except for Melby-Lervåg

and Hulme (2013), all previous meta-analyses investigating near transfer of WM training have excluded simple spans from the WM transfer domain. We decided to include simple spans as it has been argued that simple and complex spans (the latter ones being included both in previous meta-analyses and the present one) can in fact be considered as measures of the same cognitive processes (Unsworth & Engle, 2007). Nevertheless, when we ran the same analysis without the simple spans, the effect size remained similar ($g = 0.21$). Third, Jaeggi et al. (2010) have suggested that the near transfer effects may be smaller following n-back training than training with other WM training paradigms, because the n-back task shows low correlations with other WM tasks. However, a recent study showed that n-back tasks are in fact highly correlated with other WM tasks at a latent level (Schmiedek et al., 2014). Fourth, the differences between the size of the near transfer effect in the previous meta-analyses and the present one may be partly related to differences in the inclusion of studies, and the fact that we included all relevant measures from each original study in our analysis. Finally, in light of the present results, perhaps the most important explanation for higher effect sizes for WM transfer in previous meta-analyses is the fact that those studies have not separated untrained variants of the training task from other WM tasks. This could be even a more acute problem in meta-analyses on studies using several training tasks where it becomes more likely that the transfer tasks include untrained versions of the training tasks. In an attempt to investigate this, we reanalyzed the data from healthy adults included in Melby-Lervåg et al. (2016) by separating between untrained versions of the training task(s) and other untrained WM tasks. We excluded those studies in which the n-back task was the only training task in order to be able to examine if the inclusion of untrained versions of the training task(s) leads to an overestimation of the near transfer effects also for other training paradigms. The pooled effect size for all types of WM tasks was $g = 0.29$ [0.18, 0.40] $p < .001$. However, when analyzing untrained versions of the training tasks separately from other WM tasks, the results

showed that the effect size was significantly stronger, Q_M (1) = 8.89, p < .01, for the former, g = 0.62 [0.38, 0.86] p < .001 (number of effect sizes = 19), than the latter, g = 0.21, [0.08, 0.33] p < .01 (number of effect sizes = 70). The findings of this reanalysis are in line with the present meta-analysis and indicate that also for other WM training paradigms task-specific transfer is important to take into account when investigating transfer effects. These results go against the idea that the transfer effects in the present study are lower because of issues with concurrent validity of the n-back task.

For cognitive control, the transfer effect was approximately of the same size as in the two previous meta-analyses investigating transfer of different kinds of WM training to executive functions and attention (Melby-Lervåg & Hulme, 2013; Weicker et al., 2016). Also for Gf, the present transfer effect size is roughly in line with the three previous meta-analyses investigating transfer to Gf from n-back training (Au et al., 2015; Melby-Lervåg & Hulme, 2016; Melby-Lervåg et al., 2016).

In sum, in the present meta-analysis, the only notable transfer effect is seen to untrained n-back tasks. Despite the fact that the transfer effects to the other domains are also statistically significant, they can be considered very small. This is because an effect size of 0.2 means that only approximately 1% of the variance of the dependent variable (e.g., score on a Gf task) can be explained by which group (training or control group) the participant belongs to. The practical significance of such effects can thus be questioned (see also Melby-Lervåg & Hulme, 2016).

The Mechanisms behind Transfer

We hypothesized that if n-back training enhances the WM components it consists of, the magnitude of transfer effects would follow the presumed cognitive overlap between the transfer tasks and the training task (Dahlin et al., 2008; Waris et al., 2015). This would result in a gradual decrease in effect sizes, with the strongest transfer effects to untrained n-back

tasks, followed by other WM tasks, cognitive control, and Gf. However, in the present study, the only noteworthy transfer effect was seen to untrained n-back tasks, while transfer to other tasks was at similar, very small levels. This pattern of results suggests that the transfer effects of n-back training are mainly caused by acquisition of task-specific aspects such as suitable strategies, rather than better-functioning WM components. This is because task-specific improvement can enhance performance only on tasks with a similar structure where the same strategies can be successfully employed. An actual improvement in the effectiveness of the underlying WM components such as flexibility of updating and storage capacity, on the other hand, should result in broad transfer effects to different kinds of measures (von Bastian & Oberauer, 2014). A potential caveat is that a similar pattern of results as the one seen here could also be due to an improvement that is limited to the updating component. This is because we did not separate between updating tasks and other WM tasks in the analyses. However, a post hoc analysis of our data showed that the transfer effects to updating tasks other than n-back ($g = 0.26$) were roughly of the same size as the transfer effects to other WM tasks ($g = 0.23$), strengthening the conclusion that n-back training mainly improves task-specific aspects and not WM. It may thus be that the very small and similar effects of transfer to other WM tasks, Gf tasks and tasks measuring cognitive control reflect some general effects such as enhanced attention, perceptual speed, or getting used to the computer and performing demanding cognitive tasks.

Moderator Analyses

We also investigated whether the choice of control group affected the training outcome. The results showed no differences in improvement between passive and active control groups when comparing them within studies that had employed both types of control groups. Based on these results, we agree with Au et al. (2015) in that there does not seem to be any clear support for the idea that Hawthorne effects affect the results. However, when

adding the type of control group as a co-variate to our main analysis, the results revealed a small, significant effect. This could mean either of two things: the within-study comparison was underpowered, failing to find a true difference, or the training groups perform better in studies with passive controls than in studies with active controls. The latter explanation has also been proposed by Au et al. (2015).

Also, the results from the present meta-analysis do not give us reason to claim that publication bias plays a major role in the results. However, the reviewed studies contained a few observations that could be considered outliers. These were removed because they either over- or underestimated the training effect beyond what was expected from sampling error alone.

Regarding the other moderators in the present meta-analysis, the results showed no difference in transfer effects between young and old participants. On the whole, this is in line with the previous meta-analyses (Melby-Lervåg & Hulme, 2013; Schwaighofer et al., 2015; Melby-Lervåg et al., 2016). The results also showed that training with single or dual n-back tasks was equally effective in producing transfer to the four cognitive domains. This concurs with the Au et al. (2015) meta-analysis showing similar transfer effects to Gf after single and dual n-back training. The number of training hours or sessions did not affect the transfer results either, again in line with most of the previous meta-analyses. However, Weicker et al. (2016) found that the number of training sessions was positively related to the size of the transfer effect to WM. This difference in results may stem from several sources. While the present meta-analysis focused on n-back training in healthy adults, the Weicker et al. (2016) meta-analysis included studies investigating all kinds of WM training paradigms in both healthy and clinical samples of children and adults. It is in principle also possible that higher amounts of n-back training than what is currently used might yield stronger effects. Finally, in

line with the Au et al. (2015) meta-analysis, our results indicated that it did not matter if the

WM and Gf transfer tasks consisted of verbal or nonverbal material.

Limitations of the Present Study

The present meta-analysis focused on only one type of WM training, namely n-back

training. By restricting the analysis to only one training task type, it was easier to interpret the

present pattern of transfer effects. Investigating transfer elicited by several training tasks

makes it even more challenging to separate between task-specific effects and increased

effectiveness of general WM mechanisms. We were able to conclude that out of the four

cognitive domains studied here, n-back training produces substantial transfer only to

untrained variants of the training task. Transfer effects to other tasks (whether they measure

WM, Gf, or cognitive control), albeit being observable, are small and apparently of little

practical significance.

Due to the limited number of available studies, we were not able to investigate

transfer to other cognitive domains than the four studied here. For the same reason, we could

not perform all the moderator analyses for all four cognitive domains (e.g., material-specific

aspects of transfer were investigated only for WM and Gf), and we refrained from

investigating interactions between moderators.

Directions for Future Studies

There is still much controversy regarding the efficacy of WM training despite the fact

that the issue has been investigated in many training studies and a few meta-analyses. We

believe that much of the controversy is due to the great variability between training studies

regarding for example the choice of training and transfer tasks, control group, and study

population. This variability is then reflected in meta-analyses as well, because different

researchers will make different choices in categorizing tasks, participant groups, and studies.

On the one hand, one could think that variation in for example transfer tasks is important in

order to be able to draw conclusions about improvement in general WM mechanisms (for a

discussion, see e.g., Shipstead et al., 2010; Shipstead, Redick et al., 2012). This is because

cognitive tasks always involve variance that stems from other sources than the ability of

interest. Apart from random error (such as variation in alertness or disturbing noises in the

testing environment), such variance comes from other abilities engaged in solving the tasks,

strategies employed by the participants, or differences related to the type of stimuli involved.

Therefore, it has been recommended that training studies should utilize factor analysis to

analyze transfer at a latent variable level that should provide more reliable information about

the ability of interest compared to task-specific performance (Schmiedek, Lövdén, &

Lindenberger, 2010). On the other hand, the present results indicate that task-specific aspects

play an important role in the transfer effects. A failure to make a distinction between task-

specific and other WM transfer not only inflates effect sizes for near transfer, but may even

lead one to a wrong track when searching for theoretical explanations for WM training effects

(i.e., assuming that WM training increases the effectiveness of WM in general rather than

considering also alternative hypotheses on strategy-based effects). Thus, if one wants to shed

more light on the underlying mechanisms of transfer, more emphasis should be put on task-

specific aspects. It would be important to systematically analyze performances in tasks that

are closely related to the training task, in order to find out which are the mechanisms that

drive the major transfer effects.

    As previously mentioned, there is, however, not much research on what different

executive tasks actually measure and how reliable they are over time. Low task reliabilities

together with small sample sizes (often employed in training studies), result in weak statistical

power and consequently lower chances of observing a putative effect of training.

        We believe that it would be important for future studies to try to solve issues

related to task reliability and validity, pairing of training and transfer tasks, and statistical

power, rather than conducting more training studies that carry the current methodological problems. Ultimately, training effects should also be evaluated with measures that are more closely tapping real-life working memory demands. Furthermore, future studies should pave a way to a theory of the processes involved in a training-induced change in WM, a theory that is currently missing (Gibson, Gondoli, Johnson, Steeger, & Morrissey, 2012). Our results highlighting the role of task-specific transfer suggest that a considerable part of the transfer effects is related to self-generated performance strategies that emerge during a repeated practice with a limited set of WM tasks. All in all, we concur with what Shipstead, Hicks, and Engle (2012) stated already four years ago: "Working memory training remains a work in progress".

Conclusions

The present meta-analysis on the efficacy of n-back training shows medium transfer effects to untrained versions of the trained n-back tasks and small transfer effects to other WM tasks, cognitive control, and Gf. Our results thus suggest that previous meta-analyses investigating the effects of WM training have overestimated the transfer effects to WM by including untrained variants of the training tasks in their WM transfer domain. Consequently, transfer of n-back training is more task-specific than has previously been suggested.

Acknowledgement

References

*Anguera, J. A., Bernard, J. A., Jaeggi, S. M., Buschkuehl, M., Benson, B. L. Jennett, S.,

    …Seidler, R. D. (2012). The effects of working memory resource depletion and

    training on sensorimotor adaptation. *Behavioral Brain Research*, *228*(1), 107–115.

    doi:10.1016/j.bbr.2011.11.040

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015).

    Improving fluid intelligence with training on working memory: a meta-analysis.

    Psychonomic Bulletin & Review, 22(2)*, 366–377. doi:10.3758/s13423-014-0699-x

Au, J., Buschkuehl, M., Duncan, G. J., & Jaeggi, S. M. (2016). There is no convincing

    evidence that working memory training is NOT effective: A reply to Melby-Lervåg

    and Hulme (2015). *Psychonomic Bulletin & Review, 23*(1), 331–337. doi:

    10.3758/s13423-015-0967-4

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in

    Cognitive Sciences, 4*(11), 417–423. doi:10.1016/S1364-6613(00)01538-2

*Beavon, P. (2012). Improving memory using n-back training (Bachelor's thesis). Retrieved

    from http://ro.ecu.edu.au/theses_hons/65/.

Becker, B. J. (2000). Multi-variate meta-analysis. In H. E. A. Tinsley & E. D. Brown (Eds.),

    Handbook of Applied Multivariate Statistics and Mathematical Modelling (pp. 499–

    525). Orlando: Academic Press.

Bishop, D. V., & Thompson, P. A. (2016). Problems in using p-curve analysis and text-

    mining to detect rate of p-hacking and evidential value. *PeerJ*, 4:e1715.

    http://doi.org/10.7717/peerj.1715

Bogg, T. & Lasecki, L. (2014). Reliable gains? Evidence for substantially underpowered

    designs in studies of working memory training transfer to fluid intelligence. *Frontiers

    in Psychology*, *22*(5). doi: 10.3389/fpsyg.2014.01589.

Brehmer, Y., Westerberg, H., & Bäckman, L. (2012). Working-memory training in younger

and older adults: training gains, transfer, and maintenance. *Frontiers in Human*

Neuroscience, 6(63). doi:10.3389/fnhum.2012.00063

*Buschkuehl, M., Hernandez-Garcia, L., Jaeggi, S.M., Bernard, J. A., & Jonides, J. (2014).

Neural effects of short-term training on working memory. *Cognitive, Affective, &*

Behavioral Neuroscience, *14*(1), 147–160. doi: 10.3758/s13415-013-0244-9

*Bürki, C. N., Ludwig, C., Chicherio, C., & De Ribaupierre, A. (2014). Individual differences

in cognitive plasticity: an investigation of training curves in younger and older adults.

Psychological Research, 78(6), 821–835. doi:10.1007/s00426-014-0559-3

Bäckman, L., Nyberg, L., Soveri, A., Johansson, J., Andersson, M., Dahlin, E., ... Rinne, J. O.

(2011). Effects of working-memory training on striatal dopamine release. *Science*,

333(6043), 718. doi: 10.1126/science.1204978

*Chooi, W.-T., & Thompson, L. A. (2012). Working memory training does not improve

intelligence in healthy young adults. *Intelligence*, *40*(6), 531–542.

doi:10.1016/j.intell.2012.07.004

*Colom, R., Román, F. J., Abad, F. J., Shih, P. C., Privado, J., Froufe, M., … Jaeggi, S.M.

(2013). 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, 41*(5), 712–727. doi:10.1016/j.intell.2013.09.002

Dougherty, M. R., Hamovitz, T., & Tidwell, J. W. (2016). Reevaluating the effectiveness of

n-back training on transfer through the Bayesian lens: Support for the null.

Psychonomic Bulletin & Review, *23*(1), 306–316. doi: 10.3758/s13423-015-0865-9

Dahlin, E., Neely, A. S., Larsson, A., Bäckman, L., & Nyberg, L. (2008). Transfer of learning

after updating training mediated by the striatum. *Science*, *320*(5882), 1510–1512. doi:

10.1126/science.1155466

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot based method of testing

and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected

by a simple, graphical test. *BMJ*, *315*(7109), 629–634.

http://doi.org/10.1136/bmj.315.7109.629

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in

Psychological Science, 11*(1), 19–23.

Gibson, B.S., Gondoli, D.M., Johnson, A.C., Steeger, C.M., & Morrissey, R.M. (2012). The

future promise of Cogmed working memory training. *Journal of Applied Research in

Memory and Cognition, 1*(3), 214–216. doi:10.1016/j.jarmac.2012.07.003

Green, C. S., & Bavelier, D. (2008). Exercising your brain: A review of human brain

plasticity and training-induced learning. *Psychology and Aging*, *23*(4), 692–701. doi:

10.1037/a0014345

*Heffernan, A. E. (2014). The generalizability of dual n-back training in younger adults

(Master's thesis). Retrieved from http://hdl.handle.net/10222/56021

*Heinzel, S., Lorenz, R. C., Pelz, P., Heinz, A., Walter, H., Kathmann, N., … Stelzel, C.

(2016). Neural correlates of training and transfer effects in working memory in older

adults. *NeuroImage, 134*, 236–249. doi: 10.1016/j.neuroimage.2016.03.068

*Heinzel, S., Schulte, S., Onken, J., Duong, Q-L., Riemer, T. G., Heinz, A., … Rapp, M. A.

(2014). Working memory training improvements and gains in non-trained cognitive

tasks in young and older adults. *Aging, Neuropsychology, and Cognition*, *21*(2), 146–

173. doi:10.1080/13825585.2013.790338

Hindin, S. B., & Zelinski, E. M. (2012). Extended practice and aerobic exercise interventions

benefit untrained cognitive outcomes in older adults: a meta-analysis. *Journal of the

American Geriatrics Society*, *60*(1), 136–141. doi: 10.1111/j.1532-5415.2011.03761.x

*Jaeggi, S. M., Buschkuehl, M., & Jonides, J. (2009). *Working memory training and transfer*. Unpublished data presented at the ONR contractor's meeting, Arlington, VA.

*Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *PNAS, 105*(19), 6829–6833. doi:10.1073/pnas.0801268105

*Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (2014). The role of individual differences in cognitive training and transfer. *Memory and Cognition, 42*(3), 464–480. doi:10.3758/s13421-013-0364-z

*Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y.-F., Jonides, J., & Perrig, W. J. (2010). The relationship between n-back performance and matrix reasoning — implications for training and transfer. *Intelligence 38*(6), 625–635. doi:10.1016/j.intell.2010.09.001

*Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131*(1), 66–71; author reply 72–75. doi:10.1037/0033-2909.131.1.66

Karbach, J., & Verhaeghen, P. (2014). Making Working Memory Work: A Meta-Analysis of Executive-Control and Working Memory Training in Older Adults. *Psychological Science, 25*(11), 2027–2037. doi:10.1177/0956797614548725

Karr, J. E., Areshenkoff, C. N., Rast, P., & Garcia-Barrera, M. A. (2014). An empirical comparison of the therapeutic benefits of physical exercise and cognitive training on the executive functions of older adults: a meta-analysis of controlled trials. *Neuropsychology*, *28*(6), 829–845. doi: 10.1037/neu0000101

Kelly, M.E., Loughrey, D., Lawlor, B.A., Robertson, I.H., Walsh, C., & Brennan, S. (2014). The impact of cognitive training and mental stimulation on cognitive and everyday

functioning of healthy older adults: a systematic review and meta-analysis. *Ageing Research Reviews*, *15*, 28–43. doi: 10.1016/j.arr.2014.02.004

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, *55*(4), 352–358.

Klingberg, T. (2010). Training and plasticity of working memory. *Trends in Cognitive Sciences, 14*(7), 317–324. doi: 10.1016/j.tics.2010.05.002

Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlström, K., ... Westerberg, H. (2005). Computerized training of working memory in children with ADHD--a randomized, controlled trial. *Journal of the American Academy of Child and Adolescent Psychiatry, 44*(2), 177–186.

Klingberg, T., Forssberg, H., & Westerberg, H. (2002). Training of working memory in children with ADHD. *Journal of Clinical and Experimental Neuropsychology*, *24*(6), 781–791.

Kundu, B., Sutterer, D. W., Emrich, S. M., & Postle, B. R. (2013). Strengthened effective connectivity underlies transfer of working memory training to tests of short-term memory and attention. *The Journal of Neuroscience*, *33*(29), 8705–8715. doi:10.1523/JNEUROSCI.5565-12.2013

*Küper, K., & Karbach, J. (2016). Increased training complexity reduces the effectiveness of brief working memory training: evidence from short-term single and dual n-back training interventions. *Journal of Cognitive Psychology*, *28*(2), 199–208. doi: 10.1080/20445911.2015.1118106

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological Assessment*. New York: Oxford University Press.

*Lilienthal, L., Tamez, E., Shelton, J. T., Myerson, J., & Hale, S. (2013). Dual n-back training

increases the capacity of the focus of attention. *Psychonomic Bulletin and Review, 20*(1), 135–141. doi:10.3758/s13423-012-0335-6

Lövdén, M., Bäckman, L., Lindenberger, U., Schaefer, S., & Schmiedek, F. (2010). A

theoretical framework for the study of adult cognitive plasticity. *Psychological Bulletin, 136*(4), 659–676. doi: 10.1037/a0020080

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-

analytic review. *Developmental Psychology, 49*(2), 270–291. doi:10.1037/a0028228

Melby-Lervåg, M., & Hulme, C. (2016). There is no convincing evidence that working

memory training is effective: A reply to Au et al. (2014) and Karbach and Verhaeghen (2014). *Psychonomic Bulletin & Review*, *23*(1), 324–330. doi: 10.3758/s13423-015-0862-z

Melby-Lervåg, M., Redick, T., & Hulme, C. (2016). 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*, *11*(4), 512–534. doi: 10.1177/1745691616635612

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., Wager, T. D.

(2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitve Psychology*, *41*(1), 49–100.

Miller, K. M., Price, C. C., Okun, M. S., Montijo, H., & Bowers, D. (2009). Is the n-back task

a valid neuropsychological measure for assessing working memory? *Archives of Clinical Neuropsychology*, *24*(7), 711–717. doi: 10.1093/arclin/acp063

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., &

Cooper, N. J. (2009). Assessment of regression-based methods to adjust for

publication bias through a comprehensive simulation study. *BMC Medical Research Methodology*, *9*(2). doi: 10.1186/1471-2288-9-2

Morrison, A. B., & Chein, J. M. (2011). Does working memory training work? The promise and challenges of enhancing cognition by training working memory. *Psychonomic Bulletin and Review*, *18*(1), 46–60. doi: 10.3758/s13423-010-0034-0

NIMH Research Domain Criteria Workshop (2010). Working memory: Workshop proceedings. Downloaded November 8, 2016 from https://www.nimh.nih.gov/research-priorities/rdoc/working-memory-workshop-proceedings.shtml

*Oelhafen, S., Nikolaidis, A., Padovani, T., Blaser, D., Koenig, T., & Perrig, W. J. (2013). Increased parietal activity after training of interference control. *Neuropsychologia*, *51*(13), 2781–90. doi:10.1016/j.neuropsychologia.2013.08.012

Peters, J. L., Sutton, A. J., Jobes, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology*, *31*(10), 991–996. doi: 10.1016/j.jclinepi.2007.11.010

*Prandl, A. (2012). Visual memory improvement in recognition (Bachelor's thesis). Retrieved from http://ro.ecu.edu.au/theses_hons/59/.

*Preece, D. (2011). The effect of working memory (n-back) training on fluid intelligence (Bachelor's thesis). Retrieved from http://ro.ecu.edu.au/theses_hons/54/.

R Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Redick, T. S., & Lindsey, D. R. (2013). Complex span and n-back measures of working memory: a meta-analysis. *Psychonomic Bulletin & Review*, *20*(6), 1102–1113. doi: 10.3758/s13423-013-0453-9

*Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., …

  Engle, R. W. (2013). No evidence of intelligence improvement after working memory

  training: a randomized, placebo-controlled study. *Journal of Exprimental Psychology:*

  General, *142*(2), 359–379. doi:10.1037/a0029082

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). *Publication Bias in Meta-Analysis:*

  Prevention, Assessment and Adjustments. Chichester: John Wiley & Sons.

*Rudebeck, S. R., Bor, D., Ormond, A., O'Reilly, J. X. & Lee, A. C. H. (2012). A potential

  spatial working memory training task to improve both episodic memory and fluid

  intelligence. *PLoS ONE*, *7*(11): e50431. doi:10.1371/journal.pone.0050431

*Salminen, T., Frensch, P., Strobach, T., & Schubert, T. (2016). Age-specific differeneces of

  dual n-back training. *Aging, Neurospychology, and Cognition*, *23*(1), 18–39. doi:

  10.1080/13825585.2015.1031723

*Salminen, T., Strobach, T., & Schubert, T. (2012). On the impacts of working memory

  training on executive functioning. *Frontiers in Human Neuroscience*, *6*, 166.

  doi:10.3389/fnhum.2012.00166

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2010). Hundred days of cognitive training

  enhance broad cognitive abilities in adulthood: Findings from the COGITO study.

  Frontiers in Aging Neuroscience, *2*, 27. doi: 10.3389/fnagi.2010.00027

Schmiedek, F., Lövdén, M., & Lindenberger, U. (2014). A task is a task is a task: putting

  complex span, n-back, and other working memory indicators in psychometric context.

  Frontiers in Psychology, *5*. doi: 10.3389/fpsyg.2014.01475

Schwaighofer, M., Fischer, F., & Bühner, M. (2015). Does working memory training transfer?

  A meta-analysis including training conditions as moderators. *Educational*

  Psychologist, *50*(2), 138–166. doi: 10.1080/00461520.2015.1036274

*Schwarb, H., Nail, J., & Schumacher, E. H. (2015). Working memory training improves visual short-term memory capacity. *Psychological Research, 80*(1), 128–148. doi: 10.1007/s00426-015-0648-y

*Schweizer, S., Hampshire, A., & Dalgleish, T. (2011). Extending brain-training to the affective domain: Increasing cognitive and affective executive control through emotional working memory training. *PLoS ONE, 6*(9), 1–5. doi:10.1371/journal.pone.0024372

Seidler, R. D., Bernard, J. A., Buschkuehl, M., Jaeggi, S., Jonides, J., & Humfleet, J. (2010). Cognitive training as an intervention to improve driving ability in the older adult. Technical Report No. MCASTL 201001. Ann Arbor: University of Michigan.

Shipstead, Z., Hicks, K. L., & Engle, R. W. (2012). Working memory training remains a work in progress. *Journal of Applied Research in Memory and Cognition*, *1*(3), 217–219. doi:10.1016/j.jarmac.2012.07.009

Shipstead, Z., Redick, T. S., & Engle, R. W. (2010). Does working memory training generalize? *Psychologica Belgica, 50*(3 & 4), 245–276.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin, 138*(4), 628–654. doi: 10.1037/a0027473

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). p-curve and effect size. Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666–681. doi: 10.1177/1745691614553988

*Smith, S. P., Stibric, M., & Smithson, D. (2013). Exploring the effectiveness of commercial and custom-built games for cognitive training. *Computers in Human Behavior*, *29*(6), 2388–2393. doi:10.1016/j.chb.2013.05.014

Stanley, T. D. & Doucouliagos, H. (2014) Meta-regression approximations to reduce

    publication selection bias. *Research Synthesis Methods*, *5*(1), 60–78. doi:

    10.1002/jrsm.1095

*Stepankova, H., Lukavsky, J., Buschkuehl, M., Kopecek, M., Ripova, D., & Jaeggi, S. M.

    (2013). The malleability of working memory and visuospatial skills: A randomized

    controlled study in older adults. *Developmental Psychology*, *50*(4), 1049–1059. doi:

    10.1037/a0034913.

*Stephenson, C. L., & Halpern, D. F. (2013). Improved matrix reasoning is limited to training

    on tasks with a visuospatial component. *Intelligence*, *4*(5), 341–357.

    doi:10.1016/j.intell.2013.05.006

Szmalec, A., Verbruggen, F., Vandierendonck, A., & Kemps, E. (2011). Control of

    interference during working memory updating. *Journal of Experimental Psychology

    Human Perception & Performance, 37*(1), 137–151. doi: 10.1037/a0020365

*Thompson, T. W., Waskom, M. L., Garel, K.-L. A., Cardenas-Iniguez, C., Reynolds, G. O.,

    Winter, R., … Gabrieli, J. D. E. (2013). Failure of working memory training to

    enhance cognition or intelligence. *PloS ONE*, *8*(5). doi:10.1371/journal.pone.0063614

Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory:

    An examination of simple and complex span and their relation to higher order

    abilities. *Psychological Bulletin*, *133*(6), 1038–1066. doi: 0.1037/0033-

    2909.133.6.1038

*Urbánek, T., & Marček, V. (2015). Investigating the effectiveness of working memory

    training in the context of Personality Systems Interaction theory. *Psychological

    Research*, *80*(5), 877–888, doi: 10.1007/s00426-015-0687-4

van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013).

    Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*,

45(2), 576–594. http://doi.org/10.3758/s13428-012-0261-6

*Vartanian, O., Jobidon, M.-E., Bouak, F., Nakashima, A., Smith, I., Lam, Q., & Cheung, B.

(2013). Working memory training is associated with lower prefrontal cortex activation

in a divergent thinking task. *Neuroscience, 236*, 186–194. doi:

10.1016/j.neuroscience.2012.12.060

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of

Statistical Software*, *36*(3), 1–48.

Von Bastian, C. C., & Oberauer, K. (2014). Effects and mechanisms of working memory

training: a review. *Psychological Research, 78*(6), 803–820. doi:10.1007/s00426-013-

0524-6

Waris, O., Soveri; A., & Laine, M. (2015). Transfer after working memory updating training.

PLoS ONE, *10*(9). doi:10.1371/journal.pone.0138734

Weicker, J., Villringer, A., & Thöne-Otto, A. (2016). Can impaired working memory

functioning be improved by training? A meta-analysis with a special focus on brain

injured patients. *Neuropsychology*, *30*(2), 190–212. doi: 10.1037/neu0000227

*Zajac-Lamparska, L., & Trempala, J. (2016). Effects of working memory and attentional

control training and their transfer onto fluid intelligence in early and late adulthood.

Health Psychology Report, 4(1), 41–53. doi: 10.5114/hpr.2016.56846

*Zhang, Q., Buschkuehl, M., Bernat, E., & Jaeggi, S. M. (2014, May 22–25). *EEG power

changes as a function of working memory training*. Poster presented at the 26th annual

convention of the Association for Psychological Science, San Francisco, CA.

Table S1

Reason for Excluding Studies with Healthy Adults Included in Previous Meta-analyses on N-back Training

Study	Inlcuded in	Reason for exclusion	Criteria not met
Li et al. (2008)	Weicker et al (2015)	No random assignment to groups	1
Feiyue et al (2009)	Melby-Lervåg et al (2016)	Additional critical information not received	
Minear (2012)	Melby-Lervåg et al (2016)	No control group with same amount of training dose on a non-WM task	
Clouter (2013)	Au et al (2015) Melby-Lervåg et al (2016)	The task of the active control group was a dual 1-back task	
Kundu et al (2013)	Au et al (2015) Melby-Lervåg et al (2016) Weicker et al (2015)	TMS administered before training	
Minear et al (2013)	Melby-Lervåg et al (2016)	Training included several tasks	3
Nussbaumer et al (2013)	Melby-Lervåg et al (2016)	All groups trained tasks loading on WM	4
Owens et al (2013)	Weicker et al (2015)	The task of the active control group was a dual 1-back task	
Savage (2013)	Melby-Lervåg et al (2016)	The task of the active control group was a dual 1-back task	
Sprenger et al (2013)	Weicker et al (2015)	Training included several tasks	3
Horvat (2014)	Melby-Lervåg et al (2016)	Written in language not understood by any of the authors	
Onraedt et al (2014)	Weicker et al (2015)	Participants with heightened rumination	5
Katz et al (unpublished)	Au et al (2015)	Additional critical information not received	

Table S2

Age and Amount of Training for the Experimental Groups

Study	Sample	Type	*n*	Overall age			Group specific age			Age group	Total		Dose	
				M	*SD*	range	*M*	*SD*	range		Sessions	Hours	Sessions	Hours
Jaeggi et al. (2008)	1	Dual	8	24.5	4.0					Young	8	3.33	Small	Small
	2	Active[a]	7	24.5	4.0					Young				
	3	Passive	8	24.5	4.0					Young				
	4	Dual	11	25.6[b]	3.3[b]					Young	12	5.00	Small	Small
	5	Passive	11	25.6[b]	3.3[b]					Young				
	6	Dual	8	25.6[b]	3.3[b]					Young	17	7.08	Large	Large
	7	Passive	8	25.6[b]	3.3[b]					Young				
	8	Dual	7	25.6[b]	3.3[b]		25.5[b]	2.3[b]		Young	19	7.92	Large	Large
	9	Passive	8	25.6[b]	3.3[b]		25.1	1.1		Young				
Jaeggi et al. (2010)	1	Dual	25	19.4[b]	1.5[b]		19.1	1.2		Young	20	5.67	Large	Small
	2	Single	21	19.4[b]	1.5[b]		19.0	1.5		Young	20	5.67	Large	Small
	3	Passive	43	19.4[b]	1.5[b]		19.4	1.0		Young				
Schweizer et al. (2011)	1	Dual	15	25.0		21-30	25.0	2.7		Young	20	6.67	Large	Large
	2	Dual	14	25.0		21-30	25.0	2.7		Young	20	6.67	Large	Large
	3	Active	16	25.0		21-30	25.0	2.0		Young				
Preece (2012)	1	Single	29			18-66	31.7				20	5.67	Large	Small
	2	Active	27			18-66	28.2							
Salminen et al. (2012)	1	Dual	20				24.4			Young	14	4.67	Small	Small
	2	Passive	18				24.5			Young				
Chooi & Thompson (2012)	1	Dual	9	~ 20						Young	8	4.00	Small	Small
	2	Dual	15	~ 20						Young	8	4.00	Small	Small
	3	Passive	22	~ 20						Young				
	4	Dual	13	~ 20						Young	20	10.00	Large	Large
	5	Dual	11	~ 20						Young	20	10.00	Large	Large
	6	Passive	23	~ 20						Young				
Beavon &	1	Single	26			18-69	35.9				20	5.00	Large	Small

Study	Sample	Type	n	Age						Age group	Amount of training			
				Overall age			Group specific age				Total		Dose	
				M	SD	range	M	SD	range		Sessions	Hours	Sessions	Hours
Prandl (2012)	2	Active	21			18-69	40.4							
Rudebeck et al.	1	Dual	27				25.4	4.5		Young	20	6.67	Large	Large
(2012)	2	Passive	28				25.5	4.7		Young				
Anguera et al. (2012)	1	Dual	21	21.0[b]	3.0[b]					Young	21.2	8.33	Large	Large
	2	Active	22	21.0[b]	3.0[b]					Young				
Vartanian (2013)	1	Single	17	30.8	7.1	20-47			22-45	Young	3	1.00	Small	Small
	2	Active	17	30.8	7.1	20-47			20-47	Young				
Thompson et al.	1	Dual	20				21.2			Young	20	15.00	Large	Large
(2013)	2	Active	19				21.3			Young				
	3	Passive	19				23.1			Young				
Stephenson &	1	Dual	28	22.5[c]	5.8[c]					Young	20	6.67	Large	Large
Halpern (2013)	2	Single	29	22.5[c]	5.8[c]					Young	20	6.67	Large	Large
	3	Single	25	22.5[c]	5.8[c]					Young	20	6.67	Large	Large
	4	Passive	26	22.5[c]	5.8[c]					Young				
Smith et al. (2013)	1	Dual	10			18-34[c]				Young	17	5.70	Large	Small
	2	Active	10			18-34[c]				Young				
	3	Passive	9			18-34[c]				Young				
Redick et al. (2013)	1	Dual	24			18-30[b]	21.1	2.7		Young	20	10.00	Large	Large
	2	Active	29			18-30[b]	20.7	2.5		Young				
	3	Passive	20			18-30[b]	21.2	2.5		Young				
Oelhafen et al. (2013)	1	Dual	14	25.2	4.1	18-34				Young	14	5.83	Small	Small
	2	Dual	14	25.2	4.1	18-34				Young	14	5.83	Small	Small
	3	Passive	15	25.2	4.1	18-34				Young				
Colom et al. (2013)	1	Dual	28				18.4	0.9		Young	24	12.00	Large	Large
	2	Passive	28				18.2	1.2		Young				
Stepankova et al.	1	Single	20	68.1	2.6	65-74	68.2	2.6	65-74	Old	7	3.33	Small	Small
(2013)	2	Single	20	68.1	2.6	65-74	68.0	2.2	65-72	Old	18	7.50	Large	Large
	3	Passive	25	68.1	2.6	65-74	68.1	3.0	65-74	Old				
Lilienthal et al.	1	Dual	13	U						Young	8	4.00	Small	Small
(2013)	2	Dual	13	U						Young	8	4.00	Small	Small

Study	Sample	Type	n	Age						Age group	Amount of training			
				Overall age			Group specific age				Total		Dose	
				M	SD	range	M	SD	range		Sessions	Hours	Sessions	Hours
	3	Passive	26	U						Young				
Jaeggi et al. (2014)	1	Dual	25	25.2	6.5	18-45				Young	20	6.67	Large	Large
	2	Single	26	25.2	6.5	18-45				Young	20	6.67	Large	Large
	3	Active	27	25.2	6.5	18-45				Young				
Buschkuehl et al. (2014)	1	Single	21	21.8[b]	2.7[b]		22.3[b]	2.1[b]		Young	7	2.33	Small	Small
	2	Active	28	21.8[b]	2.7[b]		21.2	2.1		Young				
Bürki et al. (2014)	1	Single	22			18-38	24.7	5.3	18-38	Young	10	5.00	Small	Small
	2	Active	20			18-38	24.4	5.2	19-38	Young				
	3	Passive	21			18-38	25.5	4.5	18-35	Young				
	4	Single	22			60-84	67.6	4.7	61-81	Old	10	5.00	Small	Small
	5	Active	20			60-84	67.7	5.0	61-78	Old				
	6	Passive	23			60-84	68.6	6.5	60-84	Old				
Heinzel et al. (2014)	1	Single	15			22-30	25.9	1.9	24-30	Young	12	9.00	Small	Large
	2	Passive	15			22-30	25.6	2.1	22-30	Young				
	3	Single	15			60-75	66.1	4.7	61-75	Old	12	9.00	Small	Large
	4	Passive	15			60-75	65.6	3.9	60-72	Old				
Heffernan (2014)	1	Dual	9			19-30	22.1	2.9		Young	15	6.25	Large	Large
	2	Active	10			19-30	21.7	2.1		Young				
Schwarb et al. (2015)	1	Single	27			18-30[b]				Young	8	8.00	Small	Large
	2	Passive	25			18-30[b]				Young				
	3	Single	22			18-32[b]				Young	8	5.33	Small	Small
	4	Single	22			18-32[b]				Young	8	5.33	Small	Small
	5	Passive	22			18-32[b]				Young				
Urbánek & Marcek (2015)	1	Single	37	25.2[c]	6.5[c]	18-50[c]	25.7	6.2		Young	25	8.33	Large	Large
	2	Triple	31	25.2[c]	6.5[c]	18-50[c]	24.8	5.8		Young	25	8.33	Large	Large
	3	Active	34	25.2[c]	6.5[c]	18-50[c]	24.6	7.1		Young				
Heinzel et al. (2016)	1	Single	15			60-75[b]	66.1	4.0		Old	12	9.00	Small	Large
	2	Passive	14			60-75[b]	66.0	4.8		Old				
Salminen et al. (2016)	1	Dual	25			57-77	64.8	3.8	57-73		14	11.67	Small	Large
	2	Passive	21			57-77	65.0	3.7	59-77					

| | | | | Age | | | | | | | Amount of training | | | |
| | | | | Overall age | | | Group specific age | | | | Total | | Dose | |
Study	Sample	Type	*n*	*M*	*SD*	range	*M*	*SD*	range	Age group	Sessions	Hours	Sessions	Hours
Küper & Karbach	1	Single	18	23.3	4.1	17-36			20-36	Young	5	2.50	Small	Small
(2016)	2	Dual	18	23.3	4.1	17-36			17-26	Young	5	2.50	Small	Small
	3	Passive	18	23.3	4.1	17-36			19-32	Young				
Zajac-Lamparska &	1	Single	30			20-33	23.8	1.9	20-27	Young	10	1.75	Small	Small
Trempala (2016)	2	Active	30			20-33	23.7	3.0	20-33	Young				
	3	Passive	30			20-33	22.9	2.4	20-28	Young				
	4	Single	30			60-85	66.6	4.7	60-79	Old	10	1.75	Small	Small
	5	Active	30			60-85	68.2	7.0	60-85	Old				
	6	Passive	30			60-85	68.3	4.8	62-77	Old				
Vartanian et al.	1	Dual	14			21-50			26-45	Young	10	1.67	Small	Small
(2016)	2	Passive	14			21-50			21-50	Young				
Jaeggi et al.	1	Dual	22			18-30	21.1	1.9	18-25	Young	19	7.92	Large	Large
(unpublished)	2	Active	21			18-30	21.8	2.8	19-30	Young				
Zhang et al.	1	Single	26			19-37	23.4	3.2	19-33	Young	20	6.67	Large	Large
(unpublished)	2	Active	24			19-37	23.2	3.5	19-37	Young				
Soveri et al.	1	Dual	28	22.0	2.8	19-31	22.0	3.0	19-31	Young	15	5.00	Large	Small
(unpublished)	2	Active	32	22.0	2.8	19-31	22.1	2.6	19-29	Young				

Note. Overall age = age information for the whole sample within a study; Group specific age = age information for specific groups; Dose = dichotomized variables of amount of training (Small, Large); Sample = sample number within studies; Type = type of training or control group; Age group = dichotomized age variable (Young, Old); Dual = dual n-back training; Young = less than 60 years; Small = sessions: < 15, hours: < 6.04; Active = active control group; Passive = passive control group; Large = sessions: ≥ 15, hours: ≥ 6.04; Single = single n-back training; U = undergraduates; Old = more than 60 years; Triple = triple n-back training.

[a]Additional data not included in the published article by Jaeggi et al. (2008). [b]Before exclusion of subjects. [c]Including samples not included in the present meta-analysis.

Table S3

Effect Sizes in Hedge's g and Standard Error for the Change in Performance from Pretest to Posttest for Training and Control Groups, As Well As the Difference in the Effect Between Training and Control Groups

				Training conditions		Transfer measure			Effect sizes										
					Dose				Training groups					Control groups				Comparison	
Study	Exp	Age	Type	Sessions	Hours	Domain	Mat	Task	n	g	SE	r	A/P	n	g	SE	r	g	SE
Jaeggi et al. (2008)	1	Y	Dual	Small	Small	Gf	NV	RAPM	8	0.08	0.20	.84	A[a]	7	-0.51	0.47	.29	0.59	0.51[b]
						WM	V	Digit span forward	8	0.91	0.39	.60		7	0.36	0.20	.89	0.55	0.44
						WM	V	Digit span backward	8	0.84	0.27	.89		7	0.40	0.31	.70	0.44	0.41
						Gf	NV	RAPM					P	8	0.13	0.37	.45	-0.04	0.42
						WM	V	Digit span forward						8	0.09	0.25	.75	0.82	0.46
						WM	V	Digit span backward						8	0.54	0.31	.68	0.30	0.41
	2	Y	Dual	Small	Small	Gf	NV	BOMAT	11	0.53	0.26	.70	P	11	-0.10	0.27	.60	0.60	0.38
						WM	V	Digit span	11	0.53	0.25	.72		11	0.24	0.22	.76	0.29	0.33
						WM	V	Reading span	11	0.24	0.26	.64		11	0.71	0.36	.42	-0.47	0.44
	3	Y	Dual	Large	Large	Gf	NV	BOMAT	8	1.72	0.54	.57	P	8	0.89	0.48	.28	0.83	0.72[b]
						WM	V	Digit span	8	0.39	0.19	.89		8	0.18	0.15	.92	0.21	0.25
						WM	V	Reading span	8	0.50	0.36	.55		8	0.78	0.32	.74	-0.28	0.48
	4	Y	Dual	Large	Large	Gf	NV	BOMAT	7	1.43	0.44	.83	P	8	0.64	0.29	.77	0.79	0.53[b]
						WM	V	Digit span forward	7	0.77	0.31	.81		8	0.56	0.36	.55	0.22	0.48
						WM	V	Digit span backward	7	1.19	0.54	.33		8	-0.67	0.32	.71	1.85	0.63[b]
						WM	V	Reading span	7	1.10	0.58	.12		8	0.29	0.14	.94	0.81	0.60[b]
						WM	NV	Visuospatial span task forward[a]	7	-0.40	0.39	.51		8	0.36	0.50	.03	-0.76	0.63[b]
						WM	NV	Visuospatial span task backward[a]	7	-0.16	0.54	-.02		8	0.28	0.44	.24	-0.44	0.70[b]
						CC	V	Stroop[a]	7	0.52	0.25	.85		8	-0.02	0.31	.62	0.54	0.40
						CC	V	Task switching[a]	7	0.50	0.68	.58		8	0.07	0.45	.18	0.43	0.82[b]
Jaeggi et al. (2010)	5	Y	Dual	Large	Small	N-back		N-back	25	1.58	0.29	.58	P	41	0.24	0.12	.70	1.34	0.31
						WM	V	Operation span	25	-0.07	0.21	.43		40	0.21	0.14	.59	-0.28	0.26

Study	Exp	Age	Training conditions			Transfer measure			Effect sizes										
			Type	Dose		Domain	Mat	Task	Training groups				Control groups					Comparison	
				Sessions	Hours				*n*	*g*	*SE*	*r*	A/P	*n*	*g*	*SE*	*r*	*g*	*SE*
Jaeggi et al.						Gf	NV	RAPM	25	0.84	0.24	.43		43	0.10	0.16	.47	0.74	0.29
(2010)						Gf	NV	BOMAT	25	0.51	0.22	.48		43	0.24	0.13	.67	0.27	0.25
continued		Y	Single	Large	Small	N-back		N-back	20	1.27	0.34	.28						1.04	0.36
						WM	V	Operation span	21	-0.18	0.15	.76						-0.39	0.21
						Gf	NV	RAPM	21	0.60	0.23	.55						0.51	0.28
						Gf	NV	BOMAT	21	0.79	0.18	.80						0.55	0.22
Schweizer et al.	6	Y	Dual	Large	Large	WM	V	Digit span	15	0.86	0.24	.74	A	16	0.21	0.17	.79	0.65	0.30
(2011)						Gf	NV	RAPM	15	1.04	0.26	.77		16	-2.88	0.62	-.02	3.92	0.67[bc]
						CC		Stroop	15	1.52	0.34	.73		16	1.04	0.23	.83	0.48	0.41
		Y	Dual	Large	Large	WM	V	Digit span	14	0.61	0.22	.76						0.39	0.27
						Gf	NV	RAPM	14	0.00	0.33	.24						2.88	0.70
						CC		Stroop	14	1.70	0.38	.71						0.66	0.45
Preece	7		Single	Large	Small	Gf	NV	Figure weights	29	0.36	0.13	.79	A	27	0.15	0.13	.79	0.21	0.18
(2012)																			
Salminen et al.	8	Y	Dual	Small	Small	WM	V	Auditory-verbal	20	0.65	0.31	.12	P	18	0.53	0.26	.44	0.13	0.41
(2012)						WM	NV	Visuospatial	20	0.71	0.35	-.10		18	-0.20	0.23	.53	0.90	0.42
						WM	V	Dual-modality	20	0.36	0.27	.30		18	0.82	0.29	.39	-0.45	0.40
						CC		Dual task 1 SOA 50	20	0.33	0.13	.85		18	0.32	0.19	.69	0.01	0.24
						CC		Dual task 1 SOA 100	20	0.29	0.15	.81		18	0.39	0.21	.65	-0.10	0.25
						CC		Dual task 1 SOA 400	20	0.30	0.12	.87		18	0.22	0.15	.81	0.07	0.19
						CC		Dual task 2 SOA 50	20	0.42	0.15	.81		18	0.58	0.17	.83	-0.16	0.23
						CC		Dual task 2 SOA 100	20	0.41	0.15	.82		18	0.60	0.20	.74	-0.19	0.25
						CC		Dual task 2 SOA 400	20	0.45	0.16	.79		18	0.58	0.16	.85	-0.13	0.23
						CC		Switch trials	20	0.73	0.21	.70		18	0.52	0.26	.45	0.20	0.33
						Gf	NV	RAPM	13	-0.07	0.32	.34		9	0.62	0.19	.94	-0.68	0.37
Chooi &	9	Y	Dual	Small	Small	Gf	NV	Card rotation	9	0.46	0.29	.67	P	22	0.12	0.17	.67	0.34	0.34
Thompson						Gf	NV	Paper folding	9	0,36	0.26	.72		22	-0.00	0.16	.72	0.40	0.31
(2012)						Gf	NV	Mental rotation	9	-0.96	0.32	.78		22	0.00	0.14	.78	-0.96	0.35
						Gf	NV	RAPM	9	-0.04	0.27	.68		22	-0.50	0.18	.68	0.43	0.32
						WM	V	Operation span	9	0.14	0.24	.75		22	0.00	0.15	.75	0.14	0.28

			Training conditions			Transfer measure			Effect sizes										
				Dose					Training groups				Control groups					Comparison	
Study	Exp	Age	Type	Sessions	Hours	Domain	Mat	Task	n	g	SE	r	A/P	n	g	SE	r	g	SE
Chooi & Thompson (2012) continued						CC		Word beginning and ending	9	-0.09	0.28	.66		22	-0.50	0.19	.66	0.39	0.33
		Y	Dual	Small	Small	Gf	NV	Card rotation	15	0.28	0.22	.67						0.16	0.28
						Gf	NV	Paper folding	15	-0.17	0.20	.72						-0.12	0.25
						Gf	NV	Mental rotation	15	-0.71	0.22	.78						-0.71	0.26[c]
						Gf	NV	RAPM	15	-0.12	0.21	.68						0.34	0.28
						WM	V	Operation span	15	-0.02	0.18	.75						-0.02	0.24
						CC		Word beginning and ending	15	-0.12	0.21	.66						0.36	0.29
	10	Y	Dual	Large	Large	Gf	NV	Card rotation	13	-0.08	0.23	.67	P	23	0.46	0.18	.67	-0.53	0.29
						Gf	NV	Paper folding	13	0,21	0.21	.72		23	0.04	0.16	.72	0.17	0.26
						Gf	NV	Mental rotation	13	0.05	0.18	.78		23	-0.30	0.14	.78	0.33	0.23
						Gf	NV	RAPM	13	-0.03	0.22	.68		23	-0.20	0.17	.68	0.14	0.28
						WM	V	Operation span	13	0.09	0.20	.75		23	0.02	0.15	.75	0.08	0.25
						CC		Word beginning and ending	13	-0.51	0.25	.66		23	-0.50	0.19	.66	-0.04	0.31
		Y	Dual	Large	Large	Gf	NV	Card rotation	11	0.19	0.25	.67						-0.27	0.31
						Gf	NV	Paper folding	11	0.41	0.24	.72						0.37	0.29
						Gf	NV	Mental rotation	11	-0.05	0.20	.78						0.22	0.25
						Gf	NV	RAPM	11	-0.26	0.25	.68						-0.09	0.30
						WM	V	Operation span	11	-0.03	0.21	.75						-0.05	0.26
						CC		Word beginning and ending	11	-0.47	0.27	.66						-0.01	0.33
Beavon & Prandl (2012)	11		Single	Large	Small	WM	V	Numbers reversed	26	0.25	0.13	.80	A	21	0.23	0.19	.63	0.01	0.23
						WM	V	Auditory WM	26	0.70	0.15	.81		21	0.37	0.17	.73	0.32	0.23
						WM	NV	Picture recognition	24	0.39	0.23	.38		20	0.48	0.19	.72	-0.09	0.30
Rudebeck et al. (2012)	12	Y	Dual	Large	Large	Gf	NV	BOMAT	27	0.80	0.23	.45	P	28	0.10	0.15	.70	0.70	0.27
Anguera et al. (2012)	13	Y	Dual	Large	Large	N-back		Single 3-back	21	1.15	0.25	.65	A	22	0.52	0.19	.65	0.63	0.32

Study	Exp	Age	Training conditions			Transfer measure			Effect sizes										
			Type	Dose		Domain	Mat	Task	Training groups				Control groups					Comparison	
				Sessions	Hours				n	g	SE	r	A/P	n	g	SE	r	g	SE
Anguera et al. (2012) continued						N-back		Single 4-back	21	1.18	0.30	.42		22	0.35	0.24	.42	0.82	0.38
						WM	V	Operation span	22	0.68	0.22	.57		22	0.15	0.20	.57	0.53	0.30
						Gf	NV	Thurston's card rotation	22	0.54	0.27	.29		22	0.31	0.26	.29	0.23	0.37
Vartanian et al. (2013)	14	Y	Single	Small	Small	Gf	NV	RAPM	17	0.30	0.14	.86	A	17	-0.32	0.26	.47	0.62	0.29
Thompson et al. (2013)	15	Y	Dual	Large	Large	WM	V	Operation span	19	0.76	0.20	.77	A	19	0.32	0.16	.77	0.44	0.26
						WM	V	Reading span	20	0.29	0.10	.92		19	-0.20	0.10	.92	0.46	0.14
						Gf	NV	RAPM	20	-0.05	0.25	.40		19	-0.10	0.25	.40	0.09	0.35
						Gf	NV	Blocks	20	0.44	0.15	.82		19	0.64	0.17	.82	-0.20	0.23
						Gf	NV	Matrices	20	0.21	0.27	.27		19	0.96	0.32	.27	-0.75	0.42
						Gf	V	Similarities	20	0.69	0.27	.40		19	0.46	0.26	.40	0.23	0.38
						WM	V	Operation span					P	14	0.57	0.21	.77	0.19	0.29
						WM	V	Reading span						14	0.08	0.11	.92	0.21	0.15
						Gf	NV	RAPM						19	-0.30	0.26	.40	0.23	0.35
						Gf	NV	Blocks						19	0.54	0.16	.82	-0.10	0.22
						Gf	NV	Matrices						19	0.59	0.29	.27	-0.37	0.40
						Gf	V	Similarities						19	0.46	0.26	.40	0.23	0.38
Stephenson & Halpern (2013)	16	Y	Dual	Large	Large	Gf	NV	RAPM	28	0.62	0.13	.85	P	26	0.02	0.13	.78	0.60	0.19
						Gf	NV	Cattell's culture fair	28	0.34	0.12	.82		26	0.17	0.11	.86	0.17	0.16
						Gf	NV	Matrix reasoning (WASI)	28	0.53	0.16	.73		26	0.16	0.18	.56	0.37	0.24
						Gf	NV	Matrix reasoning (BETA-III)	28	0.86	0.24	.37		26	-0.20	0.11	.84	1.09	0.27
						Gf	NV	Mental rotation	28	0.19	0.12	.80		26	0.22	0.12	.83	-0.02	0.17
						Gf	NV	Paper folding	28	0.51	0.13	.83		26	0.00	0.18	.56	0.51	0.23
		Y	Single	Large	Large	Gf	NV	RAPM	29	0.49	0.12	.87						0.48	0.17
						Gf	NV	Cattell's culture fair	29	0.35	0.12	.84						0.18	0.16
						Gf	NV	Matrix reasoning (WASI)	29	0.55	0.15	.73						0.39	0.24

Study	Exp	Age	Training conditions			Transfer measure			Effect sizes										
			Type	Dose		Domain	Mat	Task	Training groups				Control groups					Comparison	
				Sessions	Hours				n	g	SE	r	A/P	n	g	SE	r	g	SE
Stephenson & Halpern (2013) continued						Gf	NV	Matrix reasoning (BETA-III)	29	0.47	0.18	.60						0.70	0.21
						Gf	NV	Mental rotation	29	0.13	0.11	.84						-0.09	0.16
						Gf	NV	Paper folding	29	0.44	0.11	.88						0.44	0.21
		Y	Single	Large	Large	Gf	NV	RAPM	25	0.35	0.13	.81						0.33	0.19
						Gf	NV	Cattell's culture fair	25	0.24	0.11	.85						0.07	0.16
						Gf	NV	Matrix reasoning (WASI)	25	0.36	0.17	.67						0.20	0.25
						Gf	NV	Matrix reasoning (BETA-III)	25	0.05	0.20	.48						0.29	0.23
						Gf	NV	Mental rotation	25	0.37	0.11	.89						0.16	0.16
						Gf	NV	Paper folding	25	0.16	0.15	.73						0.16	0.24
Smith et al. (2013)	17	Y	Dual	Large	Small	Gf	NV	RAPM	10	0.13	0.20	.81	A	10	0.09	0.20	.80	0.04	0.28
						Gf	NV	RAPM					P	9	0.72	0.42	.33	-0.59	0.47
Redick et al. (2013)	18	Y	Dual	Large	Large	Gf	NV	RAPM	24	-0.29	0.22	.46	A	29	-0.10	0.18	.53	-0.19	0.28
						Gf	NV	RSPM	24	-0.08	0.17	.64		29	-0.10	0.14	.74	0.00	0.22
						Gf	NV	Cattell's culture fair	24	-0.23	0.18	.61		29	0.19	0.21	.39	-0.42	0.28
						Gf	NV	Paper folding	24	0.35	0.18	.65		29	0.07	0.25	.13	0.28	0.30
						Gf	V	Letter sets	24	-0.02	0.19	.57		29	-0.50	0.23	.27	0.43	0.30
						Gf	V	Number series	24	-0.19	0.21	.49		29	-0.10	0.20	.42	-0.12	0.29
						Gf	V	Inference	24	0.23	0.19	.59		29	-0.10	0.17	.56	0.33	0.26
						Gf	V	Analogies	24	-0.40	0.26	.20		29	-0.30	0.21	.38	-0.13	0.34
						CC		SynWin	24	0.44	0.21	.49		29	0.68	0.17	.69	-0.24	0.28
						CC		Control tower	23	0.48	0.18	.67		29	0.72	0.13	.88	-0.24	0.23
						CC		ATClab	24	0.08	0.16	.68		28	0.16	0.16	.65	-0.08	0.23
						WM	NV	Symmetry span	24	0.55	0.17	.73		29	0.22	0.17	.59	0.34	0.24
						WM	V	Running letter span	24	0.34	0.14	.78		29	0.22	0.14	.73	0.11	0.20
						Gf	NV	RAPM					P	20	-0.20	0.27	.27	-0.05	0.35
						Gf	NV	RSPM						20	-0.10	0.21	.58	0.04	0.27
						Gf	NV	Cattell's culture fair						20	-0.20	0.27	.25	-0.05	0.33

			Training conditions			Transfer measure			Effect sizes										
				Dose					Training groups					Control groups				Comparison	
Study	Exp	Age	Type	Sessions	Hours	Domain	Mat	Task	n	g	SE	r	A/P	n	g	SE	r	g	SE
Redick et al.						Gf	NV	Paper folding						20	-0.00	0.26	.32	0.38	0.32
(2013)						Gf	V	Letter sets						20	-0.00	0.20	.61	0.00	0.28
continued						Gf	V	Number series						20	-0.40	0.23	.53	0.25	0.31
						Gf	V	Inference						20	0.06	0.28	.21	0.16	0.34
						Gf	V	Analogies						20	-0.60	0.19	.73	0.19	0.32
						CC		SynWin						20	0.86	0.28	.41	-0.42	0.35
						CC		Control tower						20	0.42	0.14	.84	0.06	0.23
						CC		ATClab						19	0.00	0.22	.54	0.08	0.28
						WM	NV	Symmetry span						20	0.32	0.22	.55	0.23	0.28
						WM	V	Running letter span						20	0.35	0.24	.44	-0.02	0.28
Oelhafen et al.	19	Y	Dual[d]	Small	Small	CC		ANT incongruent	14	0.80	0.19	.92	P	15	0.39	0.18	.79	0.41	0.26
(2013)						WM	V	Reading span	14	0.22	0.20	.72		15	0.33	0.34	.14	-0.11	0.40
						Gf	NV	BOMAT	14	0.68	0.38	.10		15	0.73	0.22	.77	-0.05	0.44
		Y	Dual[e]	Small	Small	CC		ANT incongruent	14	0.77	0.22	.81						0.37	0.28
						WM	V	Reading span	14	0.19	0.27	.48						-0.14	0.44
						Gf	NV	BOMAT	14	0.68	0.23	.76						-0.05	0.32
Colom et al.	20	Y	Dual	Large	Large	Gf	NV	RAPM	28	-0.01	0.17	.59	P	28	-0.30	0.16	.68	0.30	0.23
(2013)						Gf	NV	DAT-AR	28	0.63	0.20	.53		28	0.41	0.18	.61	0.22	0.27
						Gf	V	PMA-AR	28	1.03	0.20	.71		28	0.98	0.20	.68	0.05	0.28
						WM	V	Reading span	28	0.23	0.18	.57		28	0.09	0.25	.15	0.14	0.30
						WM	V	Computation span	28	0.24	0.18	.54		28	0.50	0.14	.78	-0.26	0.23
						WM	NV	Dot matrix task	28	0.88	0.18	.72		28	0.62	0.17	.67	0.26	0.25
						CC		Verbal flanker	28	0.22	0.28	-.10		28	0.21	0.21	.40	0.00	0.35
						CC		Numerical flanker	28	0.02	0.24	.21		28	0.16	0.19	.51	-0.14	0.30
						CC		Simon task	28	0.14	0.21	.38		28	-0.20	0.21	.38	0.29	0.30
Stepankova	28	O	Single	Small	Small	WM	V	Digit span	20	0.42	0.14	.84	P	25	0.25	0.14	.76	0.16	0.20
et al. (2014)						WM	V	Letter-number sequencing	20	0.42	0.18	.73		25	-0.20	0.16	.67	0.57	0.24
						Gf	NV	Block design	20	0.44	0.15	.82		25	0.33	0.13	.81	0.11	0.20
Lilienthal et al.	21	Y	Dual[f]	Small	Small	WM	V	Cued recall span task	13	0.58	0.41	.00	P	26	0.11	0.28	.00	0.48	0.49

Study	Exp	Age	Training conditions			Transfer measure			Effect sizes										
			Type	Dose		Domain	Mat	Task	Training groups				Control groups					Comparison	
				Sessions	Hours				n	g	SE	r	A/P	n	g	SE	r	g	SE
(2013)						WM	NV	Focus-switching task	13	0.50	0.40	.00		26	0.50	0.29	.00	-0.01	0.50
						CC		Focus-switching task	13	0.43	0.40	.00		26	0.53	0.29	.00	-0.10	0.49
						WM	NV	Grid span task	13	0.66	0.41	.00		26	0.36	0.28	.00	0.29	0.50
						WM	V	Operation span	13	0.17	0.39	.00		26	-0.00	0.28	.00	0.20	0.48
						WM	V	Running span task	13	0.56	0.23	.73		26	0.14	0.15	.73	0.43	0.27
		Y	Dualᵍ	Small	Small	WM	V	Cued recall span task	13	-0.03	0.39	.00						-0.14	0.48
						WM	NV	Focus-switching task	13	0,68	0.41	.00						0.17	0.50
						CC		Focus-switching task	13	0.55	0.41	.00						0.02	0.50
						WM	NV	Grid span task	13	0.73	0.42	.00						0.37	0.50
						WM	V	Operation span	13	0.02	0.39	.00						0.05	0.48
						WM	V	Running span task	13	-0.09	0.21	.73						-0.23	0.25
Jaeggi et al.	22	Y	Dual	Large	Large	Gf	NV	RAPM	25	-0.13	0.18	.61	A	27	-0.00	0.17	.60	-0.11	0.25
(2014)						Gf	NV	BOMAT	25	0.07	0.16	.68		27	-0.10	0.14	.72	0.17	0.22
						Gf	NV	Surface development	25	0.34	0.14	.80		27	0.13	0.11	.84	0.20	0.17
						Gf	NV	Space relations	25	0.25	0.15	.73		27	0.08	0.21	.43	0.16	0.26
						Gf	NV	Form board	25	0.49	0.15	.77		27	-0.00	0.14	.73	0.52	0.21
						Gf	NV	Cattell's culture fair	25	0.20	0.21	.48		27	0.00	0.20	.44	0.20	0.29
						Gf	V	ETS inferences	25	0.27	0.24	.31		27	0.38	0.17	.66	-0.11	0.29
						Gf	V	Verbal analogies	24	0.10	0.22	.40		27	0.30	0.23	.29	-0.20	0.32
		Y	Single	Large	Large	Gf	NV	RAPM	26	0.13	0.18	.59						0.16	0.25
						Gf	NV	BOMAT	26	0.46	0.22	.42						0.56	0.26
						Gf	NV	Surface development	26	0.22	0.11	.85						0.09	0.16
						Gf	NV	Space relations	26	0.44	0.14	.78						0.36	0.25
						Gf	NV	Form board	26	0.26	0.18	.61						0.29	0.23
						Gf	NV	Cattell's culture fair	26	0.18	0.21	.45						0.18	0.29
						Gf	V	ETS inferences	26	0.02	0.26	.09						-0.36	0.31
						Gf	V	Verbal analogies	26	0.00	0.24	.26						-0.30	0.33
Buschkuehl	23	Y	Single	Small	Small	N-back		Single 2-back	21	0.64	0.28	.25	A	28	0.50	0.20	.53	0.13	0.35
et al. (2014)						N-back		Single 3-back	21	1.34	0.34	.25		28	0.15	0.21	.39	1.18	0.40
						N-back		Single 4-back	21	0.74	0.25	.49		28	0.41	0.21	.42	0.33	0.33

Study	Exp	Age	Training conditions			Transfer measure			Effect sizes										
			Type	Dose		Domain	Mat	Task	Training groups				Control groups					Comparison	
				Sessions	Hours				n	g	SE	r	A/P	n	g	SE	r	g	SE
Bürki et al. (2014)	24	Y	Single	Small	Small	N-back		Single 2-back	22	0.14	0.19	.60	A	20	-0.10	0.13	.84	0.29	0.23
						WM	V	Number updating	12	0.46	0.21	.78		14	0.13	0.29	.41	0.34	0.36
						WM	V	Reading span	22	0.36	0.16	.74		20	0.27	0.17	.73	0.09	0.24
						Gf	NV	RAPM	22	0.20	0.12	.86		20	0.18	0.17	.73	0.01	0.20
						CC		Stroop	22	0.41	0.19	.63		20	0.29	0.16	.76	0.12	0.25
						N-back		Single 2-back					P	21	0.14	0.20	.58	0.00	0.28
						WM	V	Number updating						12	0.25	0.26	.62	0.21	0.33
						WM	V	Reading span						21	0.50	0.25	.39	-0.15	0.30
						Gf	NV	RAPM						21	0.21	0.15	.76	-0.02	0.19
						CC		Stroop						21	0.06	0.13	.81	0.35	0.23
	25	O	Single	Small	Small	N-back		Single 2-back	22	0.64	0.21	.63	A	20	0.24	0.11	.90	0.40	0.23
						WM	V	Number updating	12	-0.35	0.22	.75		12	0.00	0.31	.41	-0.35	0.38
						WM	V	Reading span	22	0.27	0.15	.77		20	0.14	0.16	.73	0.14	0.22
						Gf	NV	RAPM	22	0.25	0.21	.53		20	0.13	0.21	.57	0.13	0.30
						CC		Stroop	22	0.38	0.17	.74		20	0.23	0.13	.84	0.15	0.21
						N-back		Single 2-back					P	23	0.16	0.14	.78	0.48	0.25
						WM	V	Number updating						12	0.18	0.19	.80	-0.53	0.29
						WM	V	Reading span						23	0.25	0.14	.80	0.02	0.20
						Gf	NV	RAPM						23	0.34	0.15	.76	-0.09	0.26
						CC		Stroop						23	0.08	0.15	.75	0.30	0.22
Heinzel et al. (2014)	26	Y	Single	Small	Large	WM	V	Digit span forward	15	0.80	0.34	.29	P	15	0.26	0.25	.53	0.55	0.42
						WM	V	Digit span backward	15	0.54	0.21	.73		15	0.60	0.26	.60	-0.06	0.33
						Gf	NV	RSPM	15	0.67	0.27	.58		15	0.10	0.34	.13	0.58	0.43
						Gf	NV	Figural relations	14	1.88	0.46	.38		15	1.61	0.40	.47	0.27	0.61[b]
						CC		Verbal fluency	15	1.24	0.43	.01		15	0.00	0.27	.47	1.24	0.50[b]
	27	O	Single	Small	Large	WM	V	Digit span forward	15	0.21	0.18	.78	P	15	-0.60	0.20	.77	0.77	0.27
						WM	V	Digit span backward	15	0.47	0.30	.39		15	-0.10	0.21	.67	0.58	0.36
						Gf	NV	RSPM	15	0.16	0.29	.36		15	-0.00	0.25	.53	0.17	0.39
						Gf	NV	Figural relations	15	0.56	0.26	.56		15	0.23	0.20	.70	0.32	0.33
						CC		Verbal fluency	15	-0.01	0.20	.71		15	-0.10	0.18	.76	0.08	0.27

Study	Exp	Age	Training conditions			Transfer measure			Effect sizes										
			Type	Dose		Domain	Mat	Task	Training groups					Control groups				Comparison	
				Sessions	Hours				n	g	SE	r	A/P	n	g	SE	r	g	SE
Stepankova et al. (2014) continued		O	Single	Large	Large	Gf	NV	Matrix reasoning	20	0.44	0.15	.81		25	0.27	0.11	.88	0.17	0.19
						WM	V	Digit span	20	0.40	0.19	.67						0.15	0.24
						WM	V	Letter-number sequencing	20	0.40	0.20	.65						0.55	0.26
						Gf	NV	Block design	20	0.73	0.18	.81						0.40	0.22
						Gf	NV	Matrix reasoning	20	0.75	0.28	.36						0.48	0.30
Heffernan (2014)	29	Y	Dual	Large	Large	Gf	NV	Cattell's culture fair	9	1.52	0.44	.71	A	10	1.36	0.39	.71	0.16	0.59[b]
						CC		Stroop effect	9	0.22	0.28	.67		10	0.11	0.26	.67	0.11	0.38
						WM	V	Operation span	9	0.60	0.30	.70		10	0.37	0.26	.70	0.23	0.39
						WM	NV	Symmetry span	9	0.20	0.12	.94		10	0.23	0.12	.94	-0.03	0.17
						WM		Movespan task	9	0.09	0.27	.68		10	-0.50	0.27	.68	0.55	0.38
Schwarb et al. (2015)	30	Y	Single	Small	Large	WM	NV	Change detection task: numbers	26	0.44	0.19	.60	P	24	-0.02	0.23	.36	0.47	0.30
						WM	NV	Change detection task: resolution	26	0.61	0.18	.69		24	0.12	0.18	.61	0.49	0.25
						WM	V	Operation span	27	0.66	0.20	.57		25	0.01	0.15	.73	0.64	0.25
						WM	NV	Symmetry span	27	0.78	0.26	.23		25	0.38	0.15	.76	0.40	0.30
						Gf	NV	RAPM	27	0.03	0.18	.57		25	0.15	0.18	.59	-0.12	0.26
	31	Y	Single	Small	Small	WM	NV	Change detection task: numbers	22	0.75	0.27	.33	P	22	-0.22	0.24	.38	0.97	0.36
						WM	NV	Change detection task: resolution	22	0.74	0.31	.10		22	0.03	0.25	.33	0.71	0.39
						WM	V	Operation span	22	0.21	0.26	.28		21	-0.16	0.25	.37	0.38	0.36
						WM	NV	Symmetry span	22	0.69	0.24	.50		21	0.11	0.30	.03	0.58	0.39
						Gf	NV	RAPM	22	0.34	0.28	.19		22	-0.15	0.14	.78	0.49	0.31
						Gf	NV	Cattell's culture fair	22	-0.08	0.27	.20		22	-0.02	0.30	.03	-0.06	0.40
						CC		Antisaccade task	22	0.61	0.18	.74		21	0.17	0.17	.70	0.45	0.25
						CC		Flanker task	21	0.74	0.30	.20		21	0.08	0.16	.74	0.67	0.34
						WM	VNV	Short-term recall task: numbers	21	0.37	0.19	.64		22	-0.16	0.26	.28	0.54	0.32

Study	Exp	Age	Training conditions			Transfer measure			Effect sizes										
				Dose					Training groups				Control groups					Comparison	
			Type	Sessions	Hours	Domain	Mat	Task	n	g	SE	r	A/P	n	g	SE	r	g	SE
Schwarb et al. (2015)						WM	VNV	Short-term recall task: resolution	21	0.18	0.08	.94		22	-0.32	0.13	.84	0.50	0.15
conintued		Y	Single	Small	Small	WM	NV	Change detection task: numbers	22	1.05	0.25	.59						1.26	0.35
						WM	NV	Change detection task: resolution	22	0.88	0.26	.44						0.85	0.36
						WM	V	Operation span	21	0.34	0.16	.77						0.51	0.29
						WM	NV	Symmetry span	21	0.24	0.18	.69						0.13	0.35
						Gf	NV	RAPM	22	-0.08	0.19	,60						0.07	0.24
						Gf	NV	Cattell's culture fair	22	0.34	0.26	.31						0.36	0.39
						CC		Antisaccade task	21	0.76	0.20	.72						0.59	0.26
						CC		Flanker task	19	0.43	0.22	.57						0.36	0.28
						WM	VNV	Short-term recall task: numbers	21	0.29	0.21	.58						0.45	0.33
						WM	VNV	Short-term recall task: resolution	21	-0.25	0.12	.85						0.07	0.18
Urbánek & Marcek (2015)	32	Y	Single	Large	Large	Gf	NV	RAPM	37	0.05	0.17	.47	A	34	-0.30	0.14	.67	0.07	0.22
						Gf	NV	BOMAT	37	-0.06	0.15	.61		34	0.17	0.17	.49	-0.23	0.23
		Y	Triple	Large	Large	Gf	NV	RAPM	31	0.30	0.17	.55						0.33	0.22
						Gf	NV	BOMAT	31	0.18	0.18	.51						0.01	0.25
Heinzel et al. (2016)	33	O	Single	Small	Large	CC		Stroop	15	0.43	0.14	.90		14	-0.03	0.13	.87	0.47	0.20
Salminen et al. (2016)	34		Dual	Small	Large	WM	NV	Visuospatial	25	0.71	0.22	.50	P	21	0.13	0.18	.66	0.58	0.29
						WM	V	Auditory-verbal	25	0.16	0.19	.55		21	-0.10	0.24	.39	0.24	0.31
						CC		Switch trials	25	0.82	0.17	.82		21	0.54	0.20	.64	0.28	0.26
Küper & Karbach (2016)	35	Y	Single	Small	Small	N-back		3-back	18	1.12	0.28	.61	P	18	0.58	0.23	.61	0.54	0.36
						CC		Task switching	18	0.54	0.35	.00		18	0.07	0.33	.00	0.47	0.48
						CC		Stroop	18	0.50	0.34	.00		18	0.43	0.34	.00	0.07	0.48
						Gf	NV	RAPM	18	0.40	0.15	.83		18	0.19	0.14	.83	0.21	0.21
		Y	Dual	Small	Small	N-back		3-back	18	0.47	0.22	.61						-0.11	0.32

Study	Exp	Age	Type	Sessions	Hours	Domain	Mat	Task	n	g	SE	r	A/P	n	g	SE	r	g	SE
				Training conditions			Transfer measure			Effect sizes									
				Dose						Training groups				Control groups				Comparison	
Küper&						CC		Task switching	18	0.17	0.33	.00						0.11	0.47
Karbach (2016)						CC		Stroop	18	0.31	0.34	.00						-0.12	0.48
continued						Gf	NV	RAPM	18	0.35	0.15	.83						0.16	0.21
Zajac	36	Y	Single	Small	Small	Gf	NV	RAPM	30	0.32	0.10	.89	A	30	0.16	0.13	.76	0.15	0.16
-Lamparska &						Gf	NV	RAPM					P	30	0.18	0.10	.87	0.13	0.14
Trempala						CC		Attentional control	30	0.64	0.19	.58		30	0.20	0.17	.59	0.44	0.25
(2016)	37	O	Single	Small	Small	Gf	NV	RAPM	30	0.16	0.05	.96	A	30	0.46	0.11	.87	-0.30	0.12
						Gf	NV	RAPM					P	30	0.04	0.05	.96	0.12	0.07
						CC		Attentional control	30	0.74	0.15	.81		30	-0.12	0.16	.62	0.86	0.22
Vartanian et al.	38	Y	Dual	Small	Small	WM	V	Word span	14	0.44	0.20	.78	P	14	0.03	0.14	.87	0.41	0.25
(2016)						WM	NV	Matrix span	14	0.31	0.23	.69		14	0.38	0.30	.47	-0.08	0.37
						WM	NV	Visual span	14	0.59	0.27	.61		14	0.37	0.34	.28	0.22	0.44
Jaeggi et al.	39	Y	Dual	Large	Large	N-back	NV	Single n-back	21	1.06	0.23	.71	A	21	0.29	0.10	.92	0.77	0.25
(unpublished)						WM	V	Reading span	21	-0.02	0.13	.82		21	0.07	0.21	.54	-0.09	0.25
						WM	NV	Symmetry span	22	0.16	0.21	.53		21	0.54	0.15	.82	-0.38	0.26
						WM	V	Operation span	22	0.44	0.17	.75		21	0.03	0.15	.78	0.41	0.22
						WM	V	Digit span	22	0.28	0.17	.68		21	-0.28	0.17	.73	0.56	0.24
						WM	NV	Change detection	22	-0.05	0.18	.65		21	-0.35	0.19	.67	0.30	0.26
						Gf	NV	BOMAT	22	0.27	0.22	.49		21	0.34	0.22	.54	-0.06	0.31
						Gf	NV	RAPM	22	0.20	0.19	.61		21	0.18	0.18	.68	0.02	0.26
						Gf	NV	Cattell's culture fair	22	0.17	0.21	.54		21	-0.05	0.15	.75	0.22	0.26
						WM	V	Item recognition	21	0.45	0.29	.14		21	0.54	0.38	-.44	-0.09	0.48
						CC	V	Task switching	21	0.25	0.27	.24		21	0.52	0.17	.75	-0.27	0.32
Zhang et al.	40	Y	Single	Large	Large	WM	V	3-back	26	1.21	0.27	.40		23	0.09	0.16	.69	1.12	0.32
(unpublished)						WM	V	5-back	26	1.00	0.24	.48		23	0.42	0.18	.68	0.58	0.30
						Gf	NV	BOMAT	26	0.67	0.21	.54		24	0.03	0.21	.45	0.65	0.30
						Gf	NV	Surface development	26	0.36	0.14	.79		24	-0.03	0.18	.61	0.39	0.23
						Gf	NV	Space relations	26	0.45	0.17	.66		24	0.17	0.18	.60	0.29	0.25
						Gf	NV	Form board	26	0.41	0.16	.69		24	0.02	0.20	.52	0.40	0.26
						CC		CPT	26	0.29	0.17	.66		23	-0.33	0.20	.56	0.63	0.26

Study	Exp	Age	Type	Sessions	Hours	Domain	Mat	Task	n	g	SE	r	A/P	n	g	SE	r	g	SE
				Dose		Transfer measure			Training groups				Control groups					Comparison	
Soveri et al.	41	Y	Dual	Large	Small	N-back		Single n-back	23	1.06	0.28	.36	A	25	0.63	0.20	.59	0.43	0.35
(unpublished)						WM	V	Running memory	27	0.42	0.17	.63		32	0.30	0.15	.65	0.13	0.23
						WM	NV	Running memory task	19	0.17	0.32	.02		25	0.33	0.24	.29	-0.16	0.40
						WM	V	Number substitution	23	0.16	0.21	.52		30	-0.10	0.18	.52	0.22	0.27
						WM	V	Digit span forward	27	0.12	0.19	.52		32	0.24	0.19	.45	-0.12	0.27
						WM	V	Digit span backward	27	-0.06	0.17	.59		32	-0.00	0.13	.72	-0.02	0.22
						WM	NV	Corsi block forward	28	0.35	0.16	.66		31	0.13	0.19	.45	0.21	0.25
						WM	NV	Corsi block backward	28	0.12	0.20	.42		31	0.45	0.18	.56	-0.34	0.27
						CC		Simon task	23	0.13	0.18	.63		30	0.19	0.18	.53	-0.06	0.25
						CC		Number-letter task	25	0.48	0.16	.73		29	0.58	0.13	.83	-0.10	0.21
						Gf	NV	Cattell's culture fair	28	0.35	0.26	.09		32	0.17	0.25	.04	0.18	0.36

Note. Exp = experiment number; Age = dichotomization of age (young, old); Type = single, dual or triple n-back training; Sessions = dichotomization of training sessions; Hours = dichotomization of training hours; Domain = cognitive domain; Mat = task material; Task = name of task; g = Hedge's g; SE = standard error; r = Pearson correlation between pre- and posttest; A/P = type of control group; Y = young; Dual = dual n-back; Small = sessions: < 15, hours < 6.04; Gf = fluid intelligence; NV = non-verbal; RAPM = Raven's Advanced Progressive Matrices; P = passive control group; WM = working memory; V = verbal; BOMAT = Bochum Matrices Test; Large = sessions: ≥ 15, hours: ≥ 6.04; Single = single n-back; A = active control group; CC = cognitive control; SOA = stimulus onset asynchrony; WASI = Wechsler Abbreviated Scale of Intelligence; RSPM = Raven's Standard Progressive Matrices; ANT = Attention Network Test; DAT-AR = abstract reasoning subtest from the Differential Aptitude Test; PMA-R = inductive reasoning subtest from the Primary Mental Abilities Battery; O = old; Triple = triple n-back; CPT = Continuous Performance Test.

[a]Additional data not included in the published article by Jaeggi et al. (2008). [b]Training-control group comparisons that are excluded from analyses after data screening due to large SE. [c]Training-control group comparisons that are excluded from analyses after data screening due to large residuals. [d]Lure trials. [e]Non-lure trials. [f]Adaptive. [g]Non-adaptive.