Intelligence 77 (2019) 101386

Contents lists available at ScienceDirect

Intelligence

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

Working memory training does not enhance older adults' cognitive skills: A comprehensive meta-analysis

Giovanni Sala[a,*], N. Deniz Aksayli[b], K. Semir Tatlidil[c], Yasuyuki Gondo[a], Fernand Gobet[d]

[a] Graduate School of Human Sciences, Osaka University, Osaka, Japan
[b] School of Computer Science, University of Nottingham, Nottingham, United Kingdom
[c] Department of Cognitive, Linguistic & Psychological Sciences, Brown University, United States of America
[d] Department of Psychological Sciences, University of Liverpool, Liverpool, United Kingdom

ARTICLE INFO

Keywords:
WM training
Meta-analysis
Transfer
Older adults

ABSTRACT

In the last two decades, considerable efforts have been devoted to finding a way to enhance cognitive function by cognitive training. To date, the attempt to boost broad cognitive functions in the general population has failed. However, it is still possible that some cognitive training regimens exert a positive influence on specific populations, such as older adults. In this meta-analytic review, we investigated the effects of working memory (WM) training on older adults' cognitive skills. Three robust-variance-estimation meta-analyses ($N = 2140$, $m = 43$, and $k = 698$) were run to analyze the effects of the intervention on (a) the trained tasks, (b) near-transfer measures, and (c) far-transfer measures. While large effects were found for the trained tasks ($\bar{g} = 0.877$), only modest ($\bar{g} = 0.274$) and near-zero ($\bar{g} = 0.121$) effects were obtained in the near-transfer and far-transfer meta-analyses, respectively. Publication-bias analysis provided adjusted estimates that were slightly lower. Moreover, when active control groups were implemented, the far-transfer effects were null ($\bar{g} = -0.008$). Finally, the effects were highly consistent across studies (i.e., low or null true heterogeneity), especially in the near- and far-transfer models. While confirming the difficulty in obtaining transfer effects with cognitive training, these results corroborate recent empirical evidence suggesting that WM is not isomorphic with other fundamental cognitive skills such as fluid intelligence.

1. Introduction

The detrimental effects of aging on cognitive function are notorious. Cognitive skills such as executive functions, working memory, reasoning, and processing speed significantly decrease in the elderly (Salthouse, 2009). Finding a way to slow down cognitive decline or at least partially restore earlier cognitive function is a key issue for society. Working-memory (WM) training has been proposed as one possible solution to this problem.

Working memory (WM) can be defined as a cognitive system used to store and manipulate the information needed to carry out cognitive tasks (Baddeley, 1992, 2000). WM capacity, that is, the number of items WM can retain and manipulate, is strongly correlated with fluid intelligence (Engle, Tuholski, Laughlin, & Conway, 1999; Kane, Hambrick, & Conway, 2005) and several measures of cognitive control (e.g., Conway, Cowan, & Bunting, 2001; Kane & Engle, 2003; Redick, Calvo, Gay, & Engle, 2011). Furthermore, WM is correlated with reading (Peng et al., 2018) and mathematical skills (Peng, Namkung, Barnes, & Sun, 2016). WM also plays a fundamental role in cognitive development. For example, deficits in WM capacity in children are associated with attention-deficit/hyperactivity disorder (ADHD; Klingberg et al., 2005), reading and mathematical difficulties (Passolunghi, 2006; Swanson, 2006), and language impairment (Archibald & Gathercole, 2006). Given the importance of WM for a broad range of cognitive and academic skills and its strong correlation with fluid intelligence, WM training has been proposed to boost cognitive function in general (e.g., Jaeggi, Buschkuehl, Jonides, & Perrig, 2008).

The fundamental assumption behind WM training is that WM is somehow malleable to training. Several explanatory mechanisms have been hypothesized. To begin with, WM training may lead to long-lasting modifications in WM-related neural circuits that are involved in attentional control processes and fluid intelligence as well (Jaeggi et al., 2008; Klingberg, 2010). This way, fostering WM may induce benefits to other cognitive skills. This claim is also based on the fact that WM and fluid intelligence appear to have a shared capacity constraint (Halford,

* Corresponding author.
E-mail address: sala@hus.osaka-u.ac.jp (G. Sala).

https://doi.org/10.1016/j.intell.2019.101386
Received 26 June 2019; Received in revised form 10 August 2019; Accepted 4 September 2019
0160-2896/ © 2019 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/BY/4.0/).

Cowan, & Andrews, 2007). The performance in tasks measuring fluid intelligence (e.g., Raven's progressive matrices) is bounded by the amount of information that can be handled by WM. If WM capacity can be increased, then an improvement in such tasks may occur (Jaeggi et al., 2008). In turn, improving fluid intelligence would benefit many other cognitive and academic skills. Another complementary explanation of the generalization of WM training refers to the role of attentional processes in both fluid intelligence and working memory (for details, see Gray, Chabris, & Braver, 2003). Like with the other cognitive training programs, the critical assumption underlying these hypotheses is that WM training fosters domain-general mechanisms such as WM capacity and cognitive control, which in turn enhances other cognitive, academic, and real-life skills.

More recently, Taatgen (2013, 2016) has suggested that cognitive enhancement may be a byproduct of the acquisition of a particular skill. According to this hypothesis, extensive training in a given task enables individuals to acquire not only domain-specific skills (i.e., how to perform the trained task) but also small elements of more abstract production rules. These small elements do not encompass any domain-specific content and thus can be generalized across different cognitive tasks. In other words, this theory postulates the existence of domain-general skills that are combinations of processing elements. These domain-general skills can be learned by domain-specific training and transfer across cognitive tasks.

1.1. Meta-analytic evidence with younger individuals

Overall, these theoretical mechanisms have not been borne out by the empirical evidence, which has established that WM training does not exert any generalized benefit on cognitive function. Four comprehensive meta-analyses (Melby-Lervåg & Hulme, 2013; Melby-Lervåg, Redick, & Hulme, 2016; Schwaighofer, Fischer, & Buhner, 2015; Weicker, Villringer, & Thöne-Otto, 2016) have found that WM training impacts on the ability to perform the trained tasks and, to a lesser degree, tasks similar to the trained tasks (i.e., near transfer). However, WM training appears to have little or no effect on cognitive tests unrelated to the trained tasks (i.e., far transfer), especially when treated groups were compared to active control groups to rule out possible placebo effects (e.g., Melby-Lervåg et al., 2016).

Other smaller meta-analyses that were aimed at evaluating the effects of WM training in particular populations (e.g., healthy younger adults) or training regimens (e.g., *n*-back training) have confirmed this pattern of results. Au et al. (2015) carried out a meta-analysis of the effects of practicing *n*-back tasks on measures of fluid intelligence in younger adults. The re-analysis of their dataset shows that the overall effect size is around zero when experimental groups are compared to active-control groups (for a detailed discussion, see Au, Buschkuehl, Duncan, & Jaeggi, 2016; Dougherty, Hamovits, & Tidwell, 2016; Melby-Lervåg & Hulme, 2016). More recent meta-analytic evidence confirms this pattern of results. Soveri, Antfolk, Karlsson, Salo, and Laine (2017) have analyzed the impact of *n*-back training on healthy adults' cognitive skills. While a robust effect on *n*-back tasks' performance is evident, the treatment seems to exert no appreciable influence on far-transfer measures such as cognitive control or fluid intelligence, especially when active-control groups are implemented. In the same vein, Sala and Gobet (2017) have investigated the possible benefits of WM training in typically developing children and found evidence of near-transfer effects only. Similar findings have been seen in children and adolescent with learning disabilities (Aksayli, Sala, & Gobet, 2019; Peijnenborgh, Hurks, Aldenkamp, Vles, & Hendriksen, 2016). WM training thus seems to be no exception to the general difficulty of enhancing overall cognitive ability by training (e.g., Moreau, Macnamara, & Hambrick, 2018; Sala & Gobet, 2019; Simons et al., 2016).

It must be noted that most of the abovementioned empirical evidence refers to young populations such as healthy children, adolescents, and younger adults. The difficulty of boosting cognitive ability and, hence, obtaining far-transfer effects is evident in populations whose cognitive function is developing or at its full potential. By contrast, the impact of WM training is less clear with older adults' cognitive skills.

1.2. Meta-analytic evidence with older adults

The effects of WM training on older adults' cognitive skills have been the object of extensive empirical research, as can be seen by the several meta-analytic reviews that have been carried out on the topic. Below, we present a summary of the most important meta-analyses conducted so far. However, these meta-analyses are either underpowered (i.e., few studies included) or run with a suboptimal modeling approach. Thus, none of these meta-analyses has reached a definite conclusion regarding the current state of the art in this literature.

Lampit, Hallock, and Valenzuela's (2014) meta-analysis included studies about several computerized cognitive-training programs, nine of which were WM-training interventions. While small to moderate near-transfer effects were obtained following WM training, modest to null effects were found with far-transfer effects. Karbach and Verhaeghen (2014) reached more optimistic conclusions. They carried out a meta-analysis of 13 studies involving older adults and found that WM training exerted a positive impact on several near- and far-transfer cognitive measures. However, Melby-Lervåg and Hulme (2016) contested the validity of the findings. In their re-analysis, they argued that – when the intervention groups were compared to active controls and the differences at baseline controlled for – the effects were significantly smaller than the ones reported in Karbach and Verhaeghen (2014). Thus, the observed effects may have been due to placebo effects and statistical artifacts. Overall, the limited number of studies did not allow to draw any definite conclusion.

Schwaighofer et al. (2015) included 47 studies, of which ten concerned older adults. As already mentioned, this meta-analysis offered only modest evidence of far-transfer effects overall. In addition, no impact of age was observed. The effects were very small (at best, around 0.150 standardized mean difference; SMD), even without correction for publication bias, and were not sustained at follow-up (i.e., an assessment occurring several months after the end of the training). Again, the small number of the studies including older adults made it difficult to draw reliable conclusions.

In another meta-analysis, Melby-Lervåg et al. (2016) reported the same pattern of results as Melby-Lervåg and Hulme (2016). When compared to active controls, the effects of WM training were moderate on near-transfer measures and null on far-transfer measures. Nevertheless, Melby-Lervåg and colleagues' findings may have been influenced by two of their inclusion criteria. First, they included only interventions implementing computerized WM-training programs. In fact, non-computerized training regimens may even be more suitable for those older adults who are unfamiliar with technology. Second, they excluded all studies reporting near-transfer but not far-transfer measures (e.g., McAvinue et al., 2013). However, near-transfer effects may represent significant benefits for older adults. Moreover, measures of cognitive control (e.g., Stroop task) were not collected and inserted into the meta-analytic models. Such strict inclusion criteria thus limited the number of included studies ($n = 12$) and outcome measures analyzed.

Weicker et al.'s (2016) meta-analysis was more inclusive than the previous ones ($n = 22$). Both the near- and far-transfer effects were slightly larger than Melby-Lervåg and colleagues' meta-analyses. The main limitation of Weicker et al. (2016) was running many independent models including a small number of effect sizes each for different cognitive tests. Such models often lack the necessary statistical power to implement an appropriate sensitivity analysis (e.g., publication bias and outlier analysis).

Nguyen, Murphy, and Andrews (2019) focus on a variety of computerized cognitive training programs for older adults, including WM training. Their results highlight positive effects in several cognitive areas (e.g., executive functions, visuospatial skills, and processing

speed). Regarding WM training in particular, their meta-analytic review contains 21 studies and reports a medium overall effect on older adults' performance on memory tasks. Also, this meta-analysis incudes an intercept model showing a small positive overall far-transfer effect in the WM-training groups. No further analysis is conducted, probably due to the small number of effect sizes.

Importantly, none of the above meta-analyses employed multilevel modeling to correct for potential biases due to statistical dependence of the effect sizes. (It is worth noting that averaging dependent effects without applying any statistical correction can lead to biased estimates too; e.g., Cheung & Chan, 2014). Mewborn, Lindbergh, and Miller's (2017) meta-analysis examined the effects of several cognitive-training regimens on older adults' cognitive skills. This meta-analysis included 16 WM-training studies and, unlike the above meta-analyses, correctly implemented multilevel modeling. However, only a single overall effect size was computed, which represented the effects of WM training on trained tasks, near-, and far-transfer measures ($\bar{g} = 0.480$). Thus, it was not clear how much each of these measures contributed to the overall effect.

Finally, Teixeira-Santos et al.'s (2019) meta-analysis included 27 studies about WM training in healthy older adults (1130 participants). Both a multilevel model and a sensitivity analysis were implemented. So far, this meta-analysis was the largest and probably the technically soundest review on the topic. The meta-analysis was, however, somewhat limited in scope because it focused only on far-transfer effects related to reasoning. The effects of WM training on other cognitive skills such as language skills, processing speed, and executive functions were not considered, which led to the exclusion of many potentially eligible studies. The results indicated a small positive training effect on the participants' fluid intelligence and more robust, yet quite heterogeneous, effects on memory-related tasks.

1.3. The present study

The above summary highlights the fact that previous meta-analyses have not offered consistent conclusions about the potential of WM training in the elderly. We believe that these discrepancies are mainly due to three factors. First, as already mentioned, only two meta-analyses have employed multilevel modeling, none of which is a comprehensive investigation of the cognitive effects of WM training in the elderly. The way statistical dependence is (or is not) addressed may lead to different conclusions, especially when the meta-analytic models include a small number of effect sizes and clusters. Second, the categorization of potentially relevant moderators differs from meta-analysis to meta-analysis. For example, it is often not totally clear how transfer distance (near vs. far) is assigned to the effect size (for a discussion see Melby-Lervåg & Hulme, 2016). Analogously, the conditions defining active control groups are often too loose. As shown by Simons et al. (2016), active controls should be involved in cognitively engaging activities to effectively control for placebo effects. However, this more accurate criterion has never been applied in reviews about WM training in the elderly. Third, the estimation and investigation of between- and within-study variability is often inconclusive and biased. The lack of multilevel modeling or any other type of statistical correction related to the nested structure of data artificially increases heterogeneity, which leads to an inflated overall effect size when publication bias is present. Furthermore, the impact on the pre-post-test effect sizes of baseline differences between treated participants and controls has never been systematically analyzed. Not controlling for baseline differences is a notable limitation because a certain amount of noise due to regression to the mean is always present.[1] This noise can be read as true heterogeneity by the meta-analytic model. Therefore, not ruling out this bias would potentially give the illusion that the literature of interest is more mixed than it actually is. Also, it is often unclear how sampling error variance, of which true heterogeneity is a function, is calculated, which may be another source of artificial variability across and within meta-analyses. Crucially, all these statistical artifacts can bias meta-regression analysis results because the noise introduced may be interpreted as real effects due to one or more moderating variable. This state of affairs has probably led to incorrect conclusions about how (in) consistent the findings are in the literature at both the primary-study and meta-analytic level (e.g., Green et al., 2019; Pergher, Shalchy, Pahor, Jaeggi, & Seitz, 2019).

The present meta-analytic review is aimed at solving the above issues. We have built a more comprehensive series of meta-analytic models than in previous studies. In addition, this meta-analysis includes not only significantly more studies and participants than the previous ones, but also a rigorous sensitivity analysis to control for statistical dependence of the effect sizes, publication bias, and influential cases. Finally, the categorization of relevant moderators and the sampling error variance estimation have been based on more appropriate and conservative criteria.

2. Method

2.1. Literature search

In line with the PRISMA statement (Moher, Liberati, Tetzlaff, & Altman, 2009), a systematic search strategy was implemented to find the relevant studies. The methods and results are presented according to up-to-date reporting standards (Appelbaum et al., 2018). The following Boolean string was used: ("working memory training" OR "WM training") AND ("older adults" OR elderly OR seniors OR geriatrics OR ageing OR aging OR "age related"). We searched Complementary Index, Academic Search Complete, Medline, Science Direct, Psyc-Info, and ProQuest Dissertation & Theses databases to identify all the potentially relevant studies. Earlier narrative and meta-analytic reviews were examined, and their reference lists scanned. In addition, we e-mailed researchers in the field ($n = 28$) asking for unpublished/inaccessible data. We received 16 responses, seven of which were positive. Finally, we used Google Scholar to perform citation searches for three publications: Karbach and Verhaeghen (2014), Melby-Lervåg and Hulme (2013), and Melby-Lervåg et al. (2016).

2.2. Inclusion/exclusion criteria

The studies were included according to the following six criteria (selected a priori):

1. The study included an intervention aimed at training WM skills. Like in Melby-Lervåg et al. (2016), the WM tasks had to constitute at least 50% of the intervention. Training programs including tDCS were excluded. A meta-analytic review of such studies can be found in Nilsson, Lebedev, Rydström, and Lövdén (2017). Also, no study employing exergames (i.e., tasks that are both physically and cognitively demanding) was included. A meta-analysis of such studies in older adults is included in Sala et al. (2019);
2. The study included at least one control group;
3. At least one transfer measure of cognitive skill was collected. Self-reported measures (e.g., Cognitive Failure Questionnaire) were excluded;

[1] Random differences at pre-test naturally tend to disappear because participants' results at post-test naturally tend towards the mean values. The effect sizes are a function of the pre-post-test differences between the two groups. If, for example, the control group is, by chance, superior at pre-test (g_pre < 0), and no differences are found at post-test (g_post = 0), then g would be positive (because the experimental group improved more than the control group).

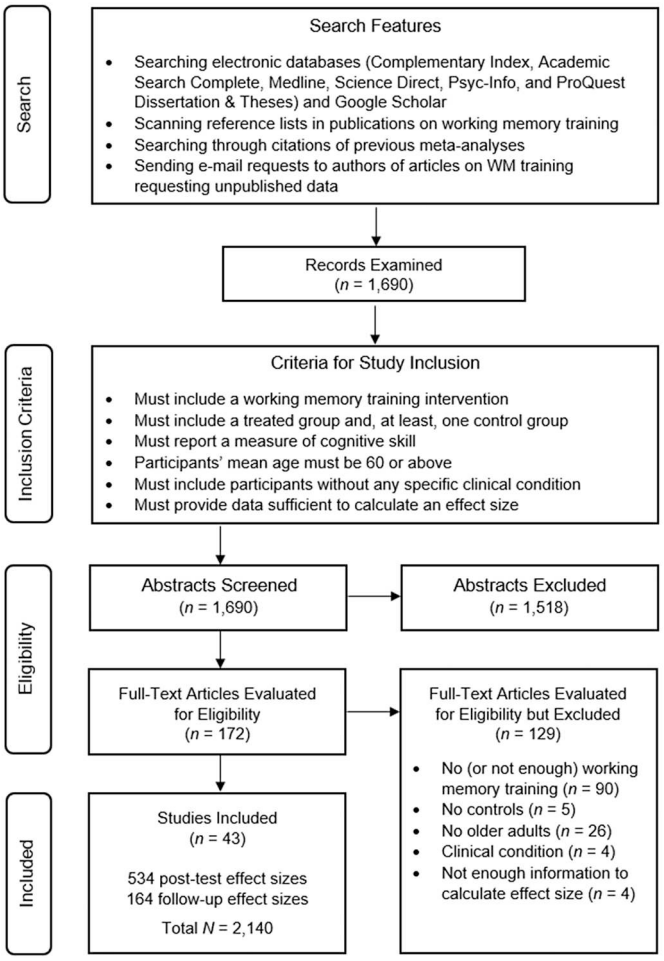

Fig. 1. Flow diagram of the search strategy in the meta-analytic review.

4. The participants' mean age was 60 or above. This criterion was consistent with a previous meta-analysis of WM training (Soveri et al., 2017);
5. The participants in the study were older adults with no serious clinical condition (e.g., dementia, brain injury);
6. The data presented in the study (or provided by the author) were sufficient to calculate an effect size.

We searched for relevant published and unpublished articles through June 1st, 2019. We found 43 studies, conducted from 2008 to 2019, that met the inclusion criteria. These studies included 534 effect sizes and a total of 2140 participants. Also, a subsample of the included studies reported follow-up effects ($k = 164$). All the data are reported in the supplemental materials available online. Fig. 1 summarizes the entire procedure.

2.3. Meta-analytic models

Consistent with Melby-Lervåg et al. (2016), each effect size was labeled as *criterion*, *near transfer*, or *far transfer*. The criterion effect sizes referred to the measures assessing the participants' performance on the trained tasks. The near-transfer effect sizes referred to memory-related measures. Examples of such measures were digit-span, *n*-back, and episodic memory tasks. Finally, far-transfer effect sizes were extracted from all the other cognitive measures. The details about the cognitive tests used in the primary studies and other descriptive statistics can be found in the Supplemental materials available online (Tables S1–S4).

Two authors coded each effect size for moderator variables independently. The Cohen's kappa was $\kappa = 0.98$. The two authors resolved every discrepancy by discussion. We ran one meta-analytic model for each type of effect size.

2.4. Moderators

We chose (a priori) four main moderators that were included in the meta-regression analyses:

1. Allocation (dichotomous variable): Whether the participants were randomly allocated to the groups; random allocation is one of the most important features to assess the quality of an intervention. Ideally, randomization ensures that the experimental and control groups do not differ with regard to any variable (e.g., cognitive skills, SES, etc.) at pre-test assessment;
2. Type of control group (active or non-active; dichotomous variable): Whether the WM training-treated group was compared to a cognitively active alternative activity (e.g., visual-search task, non-adaptive WM training). Alternative tasks with negligible cognitive demand (e.g., watching videos, physical training, and filling in questionnaires) were labeled as "non-active." This criterion was in line with commonly accepted guidelines (e.g., Boot, Simons, Stothart, & Stutts, 2013; Simons et al., 2016).[2] Inter-rater agreement was 98%;
3. Duration of training (continuous variable): The total mean time of training in hours. When the mean time of training was not provided, we used the median value of the provided range;
4. Baseline difference (continuous variable): The standardized mean difference (Hedges's *g*) between the experimental and control groups at baseline. This moderator was added to control for possible statistical artifacts (e.g., inflation of true heterogeneity) due to regression to the mean (Melby-Lervåg & Hulme, 2016). A negative and statistically significant regression coefficient would suggest the presence of some true heterogeneity due to regression to the mean (i.e., the pre-post-test gain is inversely related to the pre-test differences between the groups).

It is well-known that meta-regression often lacks adequate statistical power (Hempel et al., 2013). To minimize problems related to low statistical power (e.g., Type II error), two secondary categorical moderators (with $df > 1$) were tested separately with the Holm's method.[3] Like the Bonferroni correction, this technique allows us to adjust the *p*-value of each pairwise comparison to reduce the likelihood of Type I error (for details, see Viechtbauer, 2010):

1. Type of training task: Whether the training task employed was Cogmed, *n*-back, complex span, or other (e.g., mixed training). This categorization was the same as the one used in Melby-Lervåg et al. (2016);
2. Outcome measure: This moderator was added only in the near-transfer and far-transfer models. In the near-transfer models, the effect sizes were classified as *nearer-transfer* when the outcome

[2] Some of these control group activities were referred to as "active" in primary studies. The application of this criterion led us to recode them as non-active (e.g., Borella et al., 2014).

[3] In addition, we conducted separate analyses for the studies carried out by the researchers of the University of Padova (Borella and Carretti). This lab produced a significant number of studies about the effects of WM training on older adults' cognitive skills ($n = 10$). No other lab published more than three articles on the topic. It was thus recommendable to check for differences between Borella and Carretti's studies and studies by other researchers. This moderator was not added in the main analysis because it was confounded with the type of control group (non-active in all Borella and Carretti's studies) and duration of training (much shorter than the studies conducted by other laboratories). These analyses are reported in the supplemental materials available online.

measure was a variant of the training task (e.g., verbal *n*-back task with spatial *n*-back as the training task), *near-transfer* when the outcome measure was a common measure of STM/WM capacity, and *less-near-transfer* when the outcome measure was a proxy for episodic memory (e.g., free recall tasks).[4] The Cohen's kappa was $\kappa = 0.98$. In the far-transfer model, we categorized the effects into four groups. The effect sizes were labeled as *Gf* when referring to fluid reasoning tasks, *Gs* when referring to processing speed tasks, *EF* for executive functions tasks, and *Language* for language-related tasks (e.g., semantic comprehension). The Cohen's kappa was $\kappa = 0.96$.

This taxonomy substantially mirrors the one designed by Noack, Lövdén, Schmiedek, and Lindenberger (2009) for defining transfer distance in cognitive-training research. Based on Carroll (1993) three-stratum model, the taxonomy defines the transfer between tasks (Stratum I) from different broad skills in Stratum II as "far," the transfer between different tasks within the same Stratum II skill as "near," and the transfer between similar tasks within the same Stratum II skill as "nearest."

2.5. Effect sizes and sampling error variance calculation

The effect sizes were calculated for each relevant measure reported in the primary studies. The standardized mean difference (Cohen's *d*) was calculated with the following formula:

$$d = \frac{M_{ge} - M_{gc}}{SD_{pooled-pre}} \quad (1)$$

where M_{ge} and M_{gc} are the mean gain of the experimental group and the control group immediately after the end of the training, respectively, and $SD_{pooled\text{-}pre}$ is the pooled standard deviation of the two pre-test standard deviations. This formula represents the most appropriate way to calculate the standardized mean difference in intervention studies with a repeated-measure design (for details, see Morris, 2008; Schmidt & Hunter, 2015, pp. 352–353). When negative effects represented improved performance, the means were multiplied by -1.

We then converted the Cohen's *ds* into Hedges' *g* (Hedges & Olkin, 1985) by using the following formula

$$g = d \times \left(1 - \frac{3}{(4 \times N) - 9}\right) \quad (2)$$

where *N* is the total sample size (Schmidt & Hunter, 2015; pp. 274–275).

The formula used to calculate the sampling error variances was

$$Var_g = \left(\frac{N_e - 1}{N_e - 3} \times \left(\frac{2 \times (1 - r)}{r_{xx}} + \frac{d_e^2}{2} \times \frac{N_e}{N_e - 1}\right) \times \frac{1}{N_e} + \frac{N_c - 1}{N_c - 3} \times \left(\frac{2 \times (1 - r)}{r_{xx}} + \frac{d_c^2}{2} \times \frac{N_c}{N_c - 1}\right) \times \frac{1}{N_c}\right) \times \left(1 - \frac{3}{(4 \times N) - 9}\right)^2 \quad (3)$$

where r_{xx} is the test-retest reliability of the measure, N_e and N_c are the sizes of the experimental group and the control group, d_e and d_c are the within-group standardized mean differences of the experimental group and the control group, and *r* is the pre-post-test correlations of the experimental group and the control group, respectively (Schmidt & Hunter, 2015; pp. 343–355). Since the pre-post-test correlations and test-retest coefficients were rarely provided in the primary studies, we assumed the reliability coefficient (r_{xx}) to be equal to the pre-post-test correlation (i.e., no treatment by subject interaction was assumed; Schmidt & Hunter, 2015; pp. 350–351), and we imposed the pre-post-test correlation to be $r_{xx} = r = 0.650$.[5] This value was employed because it was the approximate mean pre-post-test correlation in Guye and von Bastian (2017), which was the largest study in the field.

2.6. Modeling approach

To prevent any bias caused by potential cherry-picking practices, we calculated the effect sizes for each relevant dependent variable reported in the studies. Several studies presented multiple-group comparisons – for example, between one experimental group and two control groups (one active and one passive), or between two experimental groups and one control group. In these cases, we calculated as many effect sizes as the number of comparisons. Our models thus included some statistically dependent effect sizes.

To control for statistical dependence between effect sizes, we used *robust variance estimation* (RVE) with hierarchical weights and small-sample corrections to calculate the overall effect size and perform meta-regression analysis (Hedges, Tipton, & Johnson, 2010). RVE allows one to model statistically dependent effect sizes and calculates adjusted (i.e., increased) overall standard errors. RVE also provides an estimation of the within-cluster true (i.e., not due to random error) heterogeneity and between-cluster true heterogeneity components (ω^2 and τ^2, respectively). We thus grouped all the effect sizes extracted from one study into the same cluster. We ran (a) intercept models to calculate the overall effect size in each meta-analytic model and (b) meta-regression models to assess the amount of true heterogeneity explained by the four main moderators. We ran the RVE models with the Robumeta software R package (Fisher, Tipton, & Zhipeng, 2017).

2.7. Sensitivity analysis

A systematic set of analyses was run to test the robustness of the results estimated by RVE. All the analyses were performed with the Metafor software R package (Viechtbauer, 2010). First, we performed Viechtbauer and Cheung's (2010) influential cases analysis in every meta-analytic model. This analysis consisted of a series of leave-one-out diagnostics for evaluating whether some effect sizes – due to their magnitude or sampling error variance – had an unusually large influence on the meta-analytic means. We excluded those influential effect sizes that contributed to the inflation of true heterogeneity. Also, we removed those effect sizes that were > 3.000 in absolute value even if they had not been detected by the influential case analysis. This criterion was applied post-hoc in order to test the robustness of the results and reduce the risk of statistical artifacts in the analyses due to inflated true heterogeneity. We report the results of the RVE models both with and without influential effect sizes.

Second, after removing the influential effect sizes, we merged the effects from the same study with the method designed by Cheung and Chan (2014). This method estimates an adjusted sampling error variance based on (a) the number of within-cluster effect sizes and (b) how homogeneous these effect sizes are (for more details, see the R codes in the supplemental materials). Then, we ran a random-effect model with the merged effect sizes. The number of reported effect sizes varied much across studies (from 1 to 34). Adopting a sample-wise procedure (i.e., merging the effects) served as a further check for the results provided by the RVE models.

Third, we ran a set of publication-bias analyses with the random-

[4] The classification of episodic-memory measures differs among meta-analyses. For example, while Weicker et al. (2016) categorize them as far-transfer, Tetlow and Edwards (2017) includes them in the near-transfer models. The latter decision, in our opinion, makes more sense. In fact, both episodic-memory tasks and WM training tasks require recall skills (for a review, see Unsworth & Engle, 2007). WM training programs may contribute to developing recall strategies such as simple mnemonics, which in turn may be used in episodic memory tasks.

[5] We replicated the analyses using different values of pre-post-test correlations ranging between 0.500 and 0.800. Only negligible differences were found.

Table 1
Overall effects in the three meta-analyses of older adults' studies sorted by significant moderators.

Model (1)	$\bar{g}$ (RVE) (2)	Adj. $\bar{g}$ (range) (3)	Heterogeneity (4)	Residual heterogeneity (5)	RE τ^2 (6)
Criterion	0.877	0.479–0.544	$\omega^2 = 0.000$, $\tau^2 = 0.252$	$\omega^2 = 0.016$, $\tau^2 = 0.019$	$\tau^2 = 0.004$ (n.s.)
Near	0.274	0.159–0.246	$\omega^2 = 0.003$, $\tau^2 = 0.033$	$\omega^2 = 0.000$, $\tau^2 = 0.000$	$\tau^2 = 0.009$ (n.s.)
Nearer	0.345	–	–	–	$\tau^2 = 0.000$ (n.s.)
Near	0.225	–	–	–	$\tau^2 = 0.006$ (n.s.)
Less-near	0.191	–	–	–	$\tau^2 = 0.000$ (n.s.)
Far	0.121	−0.030–0.113	$\omega^2 = 0.000$, $\tau^2 = 0.016$	$\omega^2 = 0.000$, $\tau^2 = 0.000$	$\tau^2 = 0.010$ (n.s.)
Non-Active	0.262	–	$\omega^2 = 0.000$, $\tau^2 = 0.040$	$\omega^2 = 0.000$, $\tau^2 = 0.010$	$\tau^2 = 0.020$ (n.s.)
Active	−0.008	–	$\omega^2 = 0.000$, $\tau^2 = 0.000$	$\omega^2 = 0.000$, $\tau^2 = 0.000$	$\tau^2 = 0.000$ (n.s.)

Note. (1) The meta-analytic model; (2) The overall RVE effect size; (3) The range of the publication bias adjusted estimates; (4) The amount of true heterogeneity of the model; (5) The heterogeneity after excluding influential cases and running meta-regression; (6) The random-effect between-study true heterogeneity after merging the statistically dependent effect sizes.

effect models (RVE does not allow to correct for publication bias analysis or test for possible influential cases). Publication bias is unanimously acknowledged as a serious problem in meta-analysis and scientific research in general (Begg & Berlin, 1988; Schmidt & Hunter, 2015; Schmidt & Oh, 2016). Therefore, it has been proposed to use multiple analyses not only to detect the possible publication bias but also to triangulate the true (i.e., unbiased) effect size (e.g., Kepes, Banks, & Oh, 2014; Kepes & McDaniel, 2015). We thus adopted a systematic and multivariate approach to assess publication bias. We produced a funnel plot depicting the relationship between the effect sizes and their standard errors. Then, we used the trim-and-fill analysis with the *L0* and *R0* estimators described in Duval and Tweedie (2000) to estimate the corrected overall effect size. The trim-and-fill analysis estimates the number of missing studies due to the systematic suppression of the null and negative effect sizes on one side of the funnel plot. The method then imputes the missing effect sizes based on the observed distribution asymmetry to generate a more symmetrical funnel plot. The adjusted overall effect size and standard error is also provided. The *L0* and *R0* estimators differ from each other regarding the type of non-parametric test they employ. Using two different estimators is recommended in order to increase the reliability of the estimates. Finally, since trim-and-fill analysis have been documented to sometimes provide false negatives (i.e., no effect sizes filled in the presence of publication bias; Simonsohn, Nelson, & Simmons, 2014), we used the PET-PEESE estimates as a further method to assess publication bias (Stanley & Doucouliagos, 2014). The PET estimator is the intercept of a weighted linear regression where the dependent variable is the effect size, the independent variable is the standard error, and the weight is the inverse of the standard error squared (i.e., precision). The PEESE estimator is obtained by replacing the standard error with the standard error squared as the independent variable. If PET suggests the presence of a real effect (i.e., intercept different from zero; $p < .100$, one-tailed), the PEESE estimator must be considered as the corrected overall effect size (Stanley, 2017; Stanley & Doucouliagos, 2014). As an updated version of the Egger's test (Egger, Smith, Schneider, & Minder, 1997), PET-PEESE can also be considered a test of symmetry of the funnel plot. It should be noted that the PET-PEESE method suffers from some shortcomings. Specifically, the technique sometimes fails to provide trustworthy results when (a) there are fewer than 20 observations, (b) true heterogeneity is high, and (c) only small-sample-size studies are present in the dataset (Stanley, 2017). For a discussion of the reliability of different publication-bias detection techniques under different conditions, see Carter, Schonbrodt, Gervais, and Hilgard (2019).

2.8. Follow-up effects

Along with data referring to immediate post-test performance, a few primary studies reported follow-up effects. The effect sizes were calculated by replacing the pre-post-test gains with the difference between the follow-up mean minus the pre-test mean in formula (1). Given the relatively small number of effect sizes in follow-up models, we did not run any sensitivity analysis.

2.9. Re-analysis of Melby-Lervåg et al. (2016)

In order to compare the effects of WM training on older adults (age ≥ 60) with younger adults (18 < age < 60), we re-analyzed a subsample of the dataset used by Melby-Lervåg et al. (2016). This dataset contained 44 studies (441 effect sizes) implementing a WM-training intervention in healthy younger adults. The categorization of the effects into three broad categories (criterion, near transfer, and far transfer) was the same as the one adopted in our meta-analytic investigation with older adults. This allowed us to make a comparison between older and younger adults' differential improvement in the three types of outcome measure. Larger gains in the population of older adults would suggest, for instance, that WM training may induce compensation effects. Furthermore, the formulas used to extract the effect sizes from primary studies was the same as the one used with older adults. For the sampling error variance, Eq. (3) could not be employed because d_e and d_c were not provided in the original dataset. We thus used the following formula (Hedges & Olkin, 1985):

$$var_g = \frac{N-1}{N-3} \times \frac{4}{N} \times \left(1 + \frac{g^2}{8}\right) \quad (4)$$

which is slightly less conservative (i.e., smaller variances on average) than Eq. (3).

We employed the same modeling approach (i.e., RVE and sensitivity analysis) as above. The dataset is included in the Supplemental materials available online.

3. Results: meta-analysis of older adults' studies

We present, in order, the criterion meta-analysis, the near-transfer meta-analysis, and the far-transfer meta-analysis. Table 1 provides a summary of the results.

3.1. Criterion meta-analysis

In this section, we examine the effects of WM training on the training tasks. Thus, this analysis does not refer to any transfer effects. Rather, it measures the impact of practicing WM training tasks on older adults' ability to perform the same tasks.

3.1.1. Main model

The RVE model included all the effect sizes related to criterion measures. The overall effect size was $\bar{g} = 0.877$, 95% CI [0.691; 1.063], $m = 28$, $k = 72$, $df = 15.61$, $p < .001$, $\omega^2 = 0.000$, $\tau^2 = 0.252$.

We ran a meta-regression model including all the four main moderators. Despite the presence of some between-cluster true heterogeneity, no moderator was significant. The type of WM training task

(secondary moderator) was not significant either (all *ps* = 1.000; Cogmed: $\overline{g}$ = 0.547, n-back: $\overline{g}$ = 0.577, complex span: $\overline{g}$ = 0.569, and other: $\overline{g}$ = 0.935).

3.1.1.1. Sensitivity analysis. One influential case was detected. Other four effect sizes were excluded because they were excessively large (g > 3.000). The results without these effect sizes were much less heterogeneous. The overall effect size was $\overline{g}$ = 0.762, 95% CI [0.613; 0.910], m = 25, k = 67, df = 14.48, p < .001, ω^2 = 0.063, τ^2 = 0.034. The meta-regression analysis showed that, this time, Baseline was a significant moderator and explained a significant amount of true heterogeneity (b = −0.528, p = .008; ω^2 = 0.016, τ^2 = 0.019). Based on Hempel et al.'s (2013) simulation, the probability of identifying a significant (p < .050) effect size of 0.200 SMD for a binary moderator in this analysis was around 40% to 50%. Also, Allocation and Type of control group approached significance (*ps* < 0.100), and thus might have been mistakenly found non-significant (Type II error). By contrast, Duration of training was largely non-significant (p = .710), which suggested that this moderator had no impact on the effect sizes regardless of statistical power. The type of WM training task (secondary moderator) was not significant (all *ps* = 1.000). Given the smaller number of studies per comparison, this moderator's power to find a significant small effect (0.200 SMD) was probably not > 20%.

3.1.1.2. Publication bias analysis. As specified in the Method section, we used Cheung and Chan's (2014) method to merge the effects (after excluding the influential cases) and perform publication bias analyses. The funnel plot is shown in Fig. 2.

The overall effect size of the random-effect model was $\overline{g}$ = 0.614, 95% CI [0.495; 0.733], p < .001, k = 25, τ^2 = 0.004. The test of heterogeneity was not significant (*QM(24)* = 24.90, p = .411). The overall effect size estimated by the trim-and-fill analysis was $\overline{g}$ = 0.544, 95% CI [0.434; 0.655], p < .001 with the *L0* estimator and $\overline{g}$ = 0.482, 95% CI [0.369; 0.596], p < .001 with the *R0* estimator. The PET and PEESE estimators were $\overline{g}$ = 0.220, 95% CI [0.085; 0.356], p = .004 and $\overline{g}$ = 0.479, 95% CI [0.392; 0.566], p < .001, respectively. In this case, the PET estimator was not reliable (p < .100, one tailed), and thus PEESE was the correct estimator.

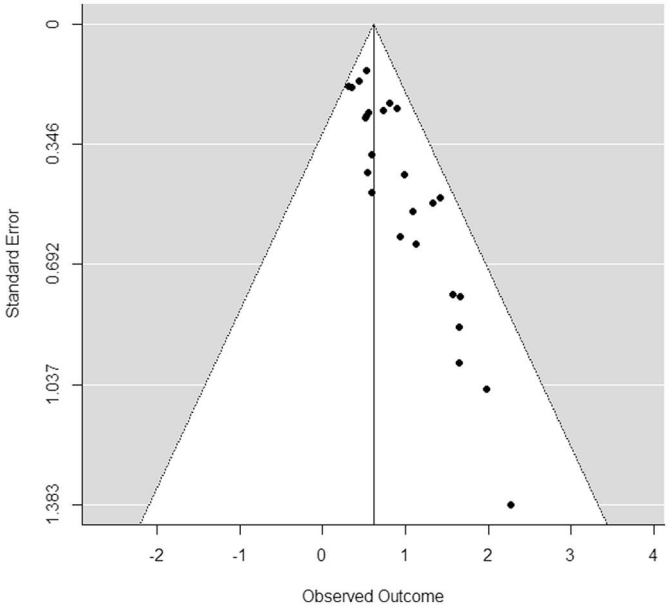

Fig. 2. Funnel plot of standard errors and effect sizes (*gs*) in the criterion meta-analysis (main model).

3.1.2. Follow-up

Ten studies reported follow-up effects. The RVE overall effect size was $\overline{g}$ = 1.016, 95% CI [0.730; 1.302], m = 10, k = 17, df = 7.20, p < .001, ω^2 = 0.035, τ^2 = 0.015.

3.1.3. Discussion

This meta-analysis analyzed the impact of the WM training programs on the trained tasks. The uncorrected overall effect size was large ($\overline{g}$ = 0.877) and relatively robust to influential case analysis and publication bias analysis. As suggested by trim-and-fill and the PEESE estimator, the unbiased effect is statistically significant and probably between $\overline{g}$ = 0.500 and $\overline{g}$ = 0.550.

The models reported some amount of true heterogeneity. A good amount of the between-cluster true heterogeneity is accounted for by a few extreme cases and baseline differences (τ^2 = 0.019). The within-clustered true heterogeneity was only partially explained by these two factors (ω^2 = 0.016). This residual true heterogeneity was probably due to the different conditions in some particular tasks (e.g., 2-back vs 0-back). In fact, the pre-test means of some task conditions were very close to the maximum value. Thus, since no substantial pre-post-test improvement was possible (i.e., ceiling effect), these effect sizes were significantly smaller than the average. This hypothesis is upheld by the fact that a small and non-significant amount of true heterogeneity is observed after merging the statistically dependent effects (τ^2 = 0.004, ns). We can thus conclude that the WM training programs did exert a meaningful and consistent impact on the participants' ability to perform the trained tasks.

3.2. Near-transfer meta-analysis

In this section, we examine the impact of WM training on the ability of older adults to perform memory tasks. These tasks are similar to the trained tasks because they tap into the same cognitive constructs (e.g., WM capacity) or the same skills (e.g., recall).

3.2.1. Main model

The RVE model included all the effect sizes related to memory measures on tasks not used during training. The overall effect size was $\overline{g}$ = 0.274, 95% CI [0.192; 0.355], m = 39, k = 214, df = 25.86, p < .001, ω^2 = 0.003, τ^2 = 0.033.

We ran a meta-regression model including all the four main moderators. Baseline difference was the only significant moderator (b = −0.416, p < .001) and explained nearly all the true heterogeneity (ω^2 = 0.000, τ^2 = 0.006). Regarding the secondary moderators, while the type of WM training task was not a significant moderator (*ps* = 1.000 in all the pairwise comparisons; Cogmed: $\overline{g}$ = 0.447 *n*-back: $\overline{g}$ = 0.176, complex span: $\overline{g}$ = 0.326, and other: $\overline{g}$ = 0.215), the similarity between trained task and outcome measure was significant. No significant difference was found between near-transfer and less-near-transfer (episodic-memory tasks) effect sizes ($\overline{g}$ = 0.225 and $\overline{g}$ = 0.191, respectively; p = .920), but the nearer-transfer effect sizes ($\overline{g}$ = 0.345) were significantly greater than near-transfer effects (p = .012) and less-near-transfer measures (p = .043). Finally, statistical power did not seem to be an issue in this case. Based on Hempel et al.'s (2013) estimations, the power (assumed g = 0.200 and alpha = 0.050) of the non-statistically significant moderators varied from about 30% (type of training task) to 80% (all the other moderators). Moreover, the residual heterogeneity was close to zero, which suggests that no other moderating variable is present beyond the ones already individuated.

3.2.1.1. Sensitivity analysis. Three influential cases were detected. The results without these effect sizes were $\overline{g}$ = 0.254, 95% CI [0.179; 0.328], m = 39, k = 211, df = 24.79, p < .001, ω^2 = 0.000, τ^2 = 0.017. Baseline difference was still the only significant moderator (b = −0.425, p < .001) and explained all the observed

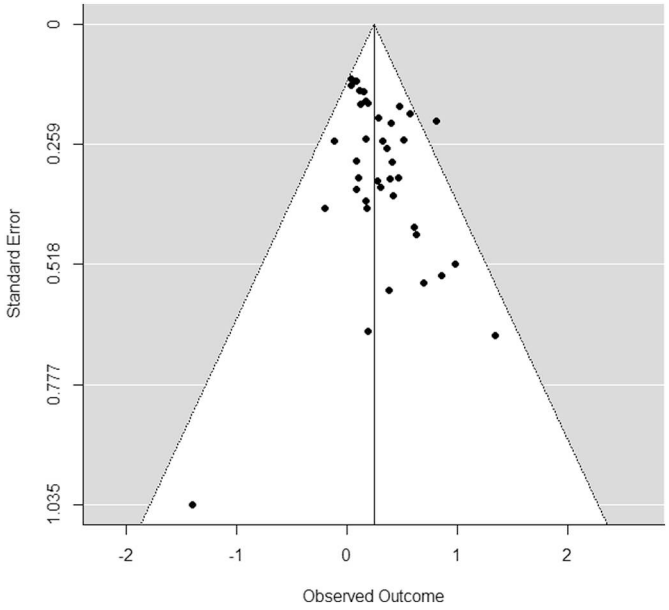

Fig. 3. Funnel plot of standard errors and effect sizes (*g*s) in the near-transfer meta-analysis (main model).

true heterogeneity ($\omega^2 = 0.000$, $\tau^2 = 0.000$). The type of WM training task was not significant (all *p*s = 1.000). Regarding the similarity between trained task and outcome measure, the only significant comparison was between nearer-transfer and near-transfer effect sizes ($p = .017$). The comparison between less-near-transfer overall effect size and nearer-transfer overall effect size showed the same trend (i.e., nearer-transfer > less-near-transfer) but, this time did not reach statistical significance ($p = .084$) due to the small number of studies in the two subgroups.

3.2.1.2. Publication bias analysis. The funnel plot is shown in Fig. 3.

With respect to publication analysis, the overall effect size of the random-effect model was $\overline{g} = 0.246$, 95% CI [0.164; 0.328], $p < .001$, $k = 39$, $\tau^2 = 0.009$. The test of heterogeneity was non-significant ($Q(38) = 38.88$, $p = .430$). The overall effect size estimated by the trim-and-fill analysis was $\overline{g} = 0.159$, 95% CI [0.067; 0.251], $p < .001$ with the *L0* estimator, and $\overline{g} = 0.246$, 95% CI [0.164; 0.328], $p < .001$ with the *R0* estimator. The PET and PEESE estimators were $\overline{g} = 0.059$, 95% CI [−0.096; 0.215], $p = .457$ and $\overline{g} = 0.189$, 95% CI [0.098; 0.280], $p < .001$, respectively. Notably, the PET estimator probably underestimated the true effect in this case because most of the effect sizes were extracted from small sample sizes (for more details, see Stanley, 2017). The other publication-bias corrected estimates (trim-and-fill and PEESE) seemed more reliable.

3.2.2. Follow-up

Fifteen studies reported follow-up effects. The RVE overall effect size was $\overline{g} = 0.378$, 95% CI [0.252; 0.504], $m = 15$, $k = 77$, $df = 9.93$, $p < .001$, $\omega^2 = 0.000$, $\tau^2 = 0.000$.

3.2.3. Discussion

This meta-analysis investigated the impact of the WM training programs on memory-related measures–measures that were thus similar, but not identical, to the trained tasks. Overall, the results showed a robust, yet modest, effect of the WM interventions on near-transfer measures. While the uncorrected overall effect was $\overline{g} = 0.274$, the sensitivity analysis estimated the probable true effect between $\overline{g} = 0.159$ and $\overline{g} = 0.246$ (both significant; as seen earlier, the PET estimator is probably overcorrected).

Regarding moderators, baseline difference, along with a few influential cases, explained all the observed true heterogeneity ($\omega^2 = 0.000$, $\tau^2 = 0.000$). The effect of the training on near-transfer outcomes was thus highly consistent across tasks and studies. As expected, the training programs tended to be more effective on nearer-transfer tasks.

Finally, the overall effect size at follow-up ($\overline{g} = 0.378$) was probably an overestimation due to selection bias. In fact, only 15 out of 39 studies (77 out of 214 effect sizes) reported follow-up measures. It is possible that, in some of the other 24 studies, follow-up measures were not collected because the effect sizes at post-test were not sufficiently large to justify further assessments.

3.3. Far-transfer meta-analysis

In this section, we examine the impact of WM training on the ability of older adults to perform non-memory-related cognitive tasks. These tasks do not share any feature with the trained tasks. That is, they tap into different skills and cognitive constructs (e.g., fluid reasoning).

3.3.1. Main model

The RVE model included all the effect sizes related to far-transfer measures. The overall effect size was $\overline{g} = 0.121$, 95% CI [0.032; 0.211], $m = 38$, $k = 248$, $df = 16.51$, $p = .011$, $\omega^2 = 0.000$, $\tau^2 = 0.016$.

We ran a meta-regression model including all the four main moderators. The type of control group and baseline differences were the only significant moderators ($b = 0.203$, $p = .009$ and $b = -0.302$, $p = .019$, respectively). This moderator explained all the observed true heterogeneity ($\omega^2 = 0.000$, $\tau^2 = 0.000$). None of the secondary moderators was significant (all *p*s ≥ 0.593; Type of training task: Cogmed: $\overline{g} = -0.054$, *n*-back: $\overline{g} = 0.092$, complex span: $\overline{g} = 0.207$, and other: $\overline{g} = 0.126$; Outcome measure: Gf: $\overline{g} = 0.095$, Gs: $\overline{g} = 0.074$, EF: $\overline{g} = 0.105$, and Lang: $\overline{g} = 0.136$). Finally, given that the residual true heterogeneity was null, it is highly improbable that other moderators could have been found to be significant or that those that were included in these analyses were incorrectly found non-significant (e.g., because of lack of statistical power).

3.3.1.1. Sensitivity analysis. Twelve influential cases were detected, none of which inflated the amount of true heterogeneity. Only one effect was excluded because it was > 3.000. The results without this effect size were $\overline{g} = 0.114$, 95% CI [0.029; 0.199], $m = 38$, $k = 247$, $df = 14.89$, $p = .012$, $\omega^2 = 0.000$, $\tau^2 = 0.004$. The only significant moderators were the type of control group and baseline differences ($p = .007$ and $p = .012$, respectively), and they explained all the observed true heterogeneity ($\omega^2 = 0.000$, $\tau^2 = 0.000$). Consequently, none of the secondary moderators was significant (all *p*s ≥ 0.434).

3.3.1.2. Publication bias analysis. The funnel plot is shown in Fig. 4.

Concerning the publication bias analysis, the overall effect size of the random-effect model was $\overline{g} = 0.114$, 95% CI [0.030; 0.199], $p = .008$, $k = 38$, $\tau^2 = 0.010$. The test of heterogeneity was non-significant ($Q(37) = 42.01$, $p = .263$). The overall effect size estimated by the trim-and-fill analysis was $\overline{g} = 0.113$, 95% CI [0.028; 0.197], $p = .009$ with the *L0* and $\overline{g} = 0.064$, 95% CI [−0.039; 0.166], $p = .221$ with the *R0* estimator. The PET and PEESE estimators were $\overline{g} = -0.030$, 95% CI [−0.172; 0.119], $p = .683$ and $\overline{g} = 0.046$, 95% CI [−0.047; 0.138], $p = .337$, respectively.

3.3.2. Type of control group

Compared to non-active controls groups, the overall effect size of the RVE model including all the effect sizes related to far-transfer measures was $\overline{g} = 0.262$, 95% CI [0.122; 0.401], $m = 25$, $k = 129$, $df = 16.60$, $p = .001$, $\omega^2 = 0.000$, $\tau^2 = 0.040$. Compared to active controls, the overall effect size was $\overline{g} = -0.008$, 95% CI [−0.105; 0.090], $m = 19$, $k = 119$, $df = 6.07$, $p = .854$, $\omega^2 = 0.000$, $\tau^2 = 0.000$.

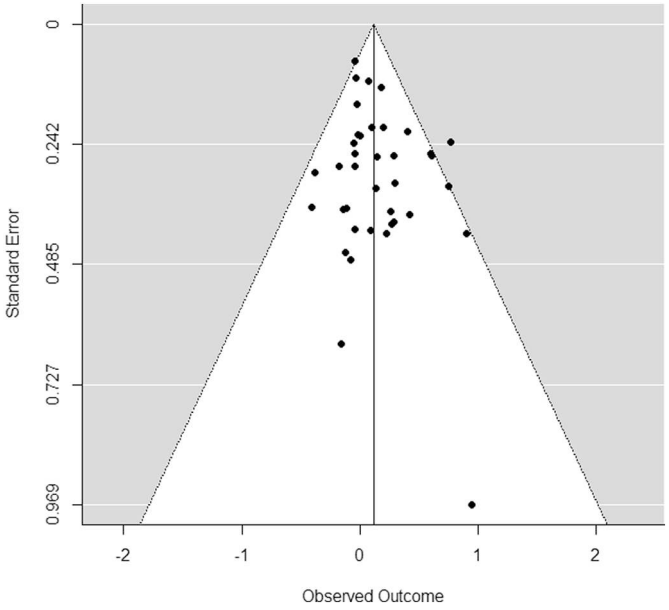

Fig. 4. Funnel plot of standard errors and effect sizes (*g*s) in the far-transfer meta-analysis (main model).

3.3.2.1. Sensitivity analysis. In the non-active-control sub-group, two influential cases were detected. The results without these effect sizes were $\overline{g}$ = 0.236, 95% CI [0.105; 0.368], m = 25, k = 127, df = 16.00, p = .002, ω^2 = 0.000, τ^2 = 0.010. In the active-control sub-group, no influential cases affecting heterogeneity were detected.

3.3.3. Follow-up

Thirteen studies reported follow-up effects. The RVE overall effect size was $\overline{g}$ = 0.241, 95% CI [0.086; 0.396], m = 13, k = 70, df = 7.92, p = .007, ω^2 = 0.000, τ^2 = 0.002.

3.3.4. Discussion

This meta-analysis investigated the impact of WM training programs on far-transfer measures – i.e., measures unrelated to the trained tasks. Overall, the analyses showed modest to null overall effect sizes that were highly homogeneous. The only significant moderators were type of control group and baseline difference. While a modest positive effect ($\overline{g}$ = 0.262) was observed in studies implementing non-active control groups, when active-control groups were employed the effect was practically zero ($\overline{g}$ = −0.008). Finally, like in the near-transfer meta-analysis, the proportion of studies reporting follow-up measures was small (13 out of 38, 70 out of 248 effect sizes), which made the estimated overall effect ($\overline{g}$ = 0.241) unreliable.

4. *Re*-analysis of Melby-Lervåg et al. (2016)

We reanalyzed the data reported in Melby-Lervåg et al. (2016) concerning the effects of WM training on younger adults. As specified above, this analysis allowed us to make a direct comparison between the effects of WM training on younger adults' and older adults' cognitive skills. The categorization of the effects (criterion, near transfer, and far transfer) and the modeling approach was the same as in the models examining WM training in the older adults. Table 2 summarizes the main results.

4.1. Criterion meta-analysis

4.1.1. Main model

The RVE model included all the effect sizes related to criterion measures. The overall effect size was $\overline{g}$ = 1.170, 95% CI [0.713; 1.627], m = 12, k = 34, df = 5.99, p < .001, ω^2 = 0.061, τ^2 = 0.392.

4.1.1.1. Sensitivity analysis. Two influential cases were detected. Another effect size was excluded because it was excessively large. The results without these effect sizes were $\overline{g}$ = 0.958, 95% CI [0.660; 1.256], m = 10, k = 31, df = 4.94, p < .001, ω^2 = 0.020, τ^2 = 0.142.

4.1.1.2. Publication bias analysis. After merging the effects, the overall effect size of the random-effect model was $\overline{g}$ = 0.836, 95% CI [0.624; 1.049], p < .001, k = 10, τ^2 = 0.039. The test of heterogeneity was non-significant ($Q(9)$ = 15.16, p = .087). The overall effect size estimated by the trim-and-fill analysis was $\overline{g}$ = 0.823, 95% CI [0.609; 1.036], p < .001 with the *L0* estimator; no bias was detected with the *R0* estimator. The PET and PEESE estimators were $\overline{g}$ = 0.841, 95% CI [0.396; 1.286], p = .006 and $\overline{g}$ = 0.821, 95% CI [0.598; 1.043], p < .001, respectively.

4.1.2. Discussion

This meta-analysis analyzed the impact of the WM training programs on the trained tasks. The unadjusted overall effect size was large ($\overline{g}$ = 1.170) and heterogeneous. The exclusion of some extreme effects and merging of the statistically dependent effect sizes reduced the true heterogeneity to a non-significant amount (τ^2 = 0.039, p = .087). Publication bias analysis estimated a smaller, yet still large, overall effect size ($\overline{g}$ about 0.820). These effects were systematically greater than the ones in the older adults.

4.2. Near-transfer meta-analysis

4.2.1. Main model

The RVE model included all the effect sizes related to near-transfer measures. The overall effect size was $\overline{g}$ = 0.208, 95% CI [0.129; 0.286], m = 31, k = 160, df = 11.82, p < .001, ω^2 = 0.053, τ^2 = 0.008.

4.2.1.1. Sensitivity analysis. Three influential cases were detected. The results without these effect sizes were $\overline{g}$ = 0.183, 95% CI [0.105; 0.260], m = 31, k = 157, df = 11.39, p < .001, ω^2 = 0.023, τ^2 = 0.013.

4.2.1.2. Publication bias analysis. After merging the effects, the overall effect size of the random-effect model was $\overline{g}$ = 0.180, 95% CI [0.119; 0.241], p < .001, k = 31, τ^2 = 0.000. The overall effect size estimated by the trim-and-fill analysis was $\overline{g}$ = 0.176, 95% CI [0.115; 0.237], p < .001 with the *L0* and $\overline{g}$ = 0.162, 95% CI [0.090; 0.233], p < .001 with the *R0* estimator. The PET and PEESE estimators were $\overline{g}$ = 0.165, 95% CI [0.060; 0.269], p = .004 and $\overline{g}$ = 0.169, 95% CI [0.095; 0.242], p < .001, respectively.

4.2.2. Discussion

This meta-analysis examined the impact of the WM training programs on memory-related measures. Overall, the results showed a slightly smaller effect than the one estimated in older adults ($\overline{g}$ = 0.208 vs. $\overline{g}$ = 0.274). The adjusted estimates ranged between $\overline{g}$ = 0.162 and $\overline{g}$ = 0.176. Like with older adults, the effects were substantially homogenous, with most of the true heterogeneity explained by a few effect sizes. No true heterogeneity was observed after merging the statistically dependent effect sizes.

4.3. Far-transfer meta-analysis

4.3.1. Main model

The RVE model included all the effect sizes related to far-transfer measures. The overall effect size was $\overline{g}$ = 0.099, 95% CI [0.015; 0.182], m = 44, k = 247, df = 15.26, p = .024, ω^2 = 0.000, τ^2 < 0.001.

4.3.1.1. Sensitivity analysis. Six influential cases were detected, three of

Table 2
Overall effects in the re-analysis of melby-lervåg et al. sorted by significant moderators.

Model (1)	$\overline{g}$ (RVE) (2)	Adj. $\overline{g}$ (range) (3)	Heterogeneity (4)	Residual heterogeneity (5)	RE τ^2 (6)
Criterion	1.170	0.821–0.823	$\omega^2 = 0.061$, $\tau^2 = 0.392$	$\omega^2 = 0.020$, $\tau^2 = 0.142$	$\tau^2 = 0.039$ (n.s.)
Near	0.208	0.162–0.176	$\omega^2 = 0.053$, $\tau^2 = 0.008$	$\omega^2 = 0.023$, $\tau^2 = 0.013$	$\tau^2 = 0.000$ (n.s.)
Far	0.099	−0.003–0.105	$\omega^2 = 0.000$, $\tau^2 < 0.001$	$\omega^2 = 0.000$, $\tau^2 = 0.000$	$\tau^2 = 0.000$ (n.s.)
Non-Active	0.161	–	$\omega^2 = 0.000$, $\tau^2 = 0.016$	$\omega^2 = 0.000$, $\tau^2 = 0.000$	$\tau^2 = 0.006$ (n.s.)
Active	0.059	–	$\omega^2 = 0.000$, $\tau^2 = 0.000$	$\omega^2 = 0.000$, $\tau^2 = 0.000$	$\tau^2 = 0.000$ (n.s.)

Note. See Note to Table 1 for abbreviations.

which were excluded because they inflated the amount of true heterogeneity. The results without these effect sizes were $\overline{g} = 0.097$, 95% CI [0.024; 0.170], $m = 44$, $k = 244$, $df = 15.96$, $p = .012$, $\omega^2 = 0.000$, $\tau^2 = 0.000$.

4.3.1.2. Publication bias analysis. After merging the effects, the overall effect size of the random-effect model was $\overline{g} = 0.107$, 95% CI [0.043; 0.170], $p < .001$, $k = 44$, $\tau^2 = 0.000$. The overall effect size estimated by the trim-and-fill analysis was $\overline{g} = 0.092$, 95% CI [0.029; 0.155], $p = .004$ with the *L0* and $\overline{g} = 0.105$, 95% CI [0.042; 0.168], $p < .001$ with the *R0* estimator. The PET and PEESE estimators were $\overline{g} = -0.003$, 95% CI [−0.119; 0.113], $p = .961$ and $\overline{g} = 0.061$, 95% CI [−0.003; 0.126], $p = .069$, respectively.

4.3.2. Type of control group

The overall effect size of the RVE model including all the effect sizes related to far-transfer measures in groups compared to non-active controls was $\overline{g} = 0.161$, 95% CI [0.035; 0.288], $m = 26$, $k = 113$, $df = 8.62$, $p = .018$, $\omega^2 = 0.000$, $\tau^2 = 0.016$. The overall effect size of the RVE model including all the effect sizes related to far-transfer measures in groups compared to active controls was $\overline{g} = 0.059$, 95% CI [−0.030; 0.147], $m = 27$, $k = 134$, $df = 9.91$, $p = .170$, $\omega^2 = 0.000$, $\tau^2 = 0.000$.

4.3.2.1. Sensitivity analysis. In the non-active-control sub-group, the results without the three influential effect sizes were $\overline{g} = 0.138$, 95% CI [0.020; 0.256], $m = 26$, $k = 110$, $df = 7.98$, $p = .027$, $\omega^2 = 0.000$, $\tau^2 = 0.000$. None of the three influential cases was included in the active-control subgroup.

4.3.3. Discussion

This meta-analysis investigated the impact of WM training programs on far-transfer measures – i.e., measures unrelated to the trained tasks. Like in the meta-analysis of WM training in older adults, the analyses showed highly homogeneous small (with non-active controls) to near-zero (with active controls) overall effect sizes.

5. General discussion

This meta-analytic investigation has addressed the question of the impact of WM training on older adults' cognitive skills. The three meta-analyses provide a picture consistent with the literature on WM training in children and young adults: strong effects in the trained tasks, small effects (near transfer and episodic memory) to medium effects (nearer transfer) in the memory tasks, and small (with non-active controls) to null effects (with active controls) in the far-transfer tasks. Table 1 summarizes the main findings.

Crucially, the meta-analytic models, especially the ones examining transfer effects, exhibit high consistency both within-study and between-study (i.e., very small or null ω^2 and τ^2). In all three meta-analyses, most of (or all) the residual true heterogeneity (if any) was accounted for by a few influential cases, differences at baseline, and type of control group. No other moderator was significant. Put together, these outcomes show that the results reported by the studies assessing the impact of WM training on older adults' cognitive function are actually extremely consistent. Therefore, there is no reason to think that this literature has produced mixed results so far.

The re-analysis of Melby-Lervåg et al.'s (2016) dataset on younger adults yielded similar outcomes: robust criterion effects, small near-transfer effects, and near-zero far-transfer effects (Table 2).

Consistent with the analysis of older adults' results, most of the models showed low true heterogeneity, which is mostly due to a small number of influential cases. The only notable difference between younger and older adults was the size of the overall criterion effects. While the corrected overall effect ranged between 0.500 and 0.550 SMD in the older adults, the estimate was about 0.800–0.850 SMD with younger adults. This difference is probably due to younger populations being better at acquiring new skills by training. Nonetheless, no appreciable difference was observed with regard to transfer effects.

Finally, it must be noted that most of the primary studies do not report any information about the reliability coefficients of the tests used. Therefore, no correction for measurement error has been applied in any of the meta-analytic models. Although this objectively constitutes a technical limitation, we think that its practical consequences are minimal, especially for what concerns near- and far-transfer models. In fact, applying a multiplicative correction to such small or null effect sizes would result in adjustments of a few hundredths of standardized mean difference at best.

5.1. Theoretical and practical implications

Our findings suggest that WM training in older adults represents no exception to the general difficulty of enhancing overall cognitive ability by training. The size of the effects was directly related to the overlap between the outcome measures and the trained tasks. More generally, the findings further establish that training leads to only limited generalization across different domains of skills. In fact, these findings echo those obtained with other cognitive-training regimens such as video-game training and brain-training programs (e.g., Rebok et al., 2014; Sala et al., 2019; Simons et al., 2016). The vast amount of negative evidence acquired so far leads us to conclude that even assuming that cognitive skills are trainable, the benefits remain domain specific. These results are in line with the prediction of theories of skill acquisition based on task domain-specificity (Chase & Ericsson, 1982; Gathercole, Dunning, Holmes, & Norris, 2019; Gobet, 2016; Gobet & Simon, 1996). Conversely, those theories predicting far transfer are not supported (e.g., Jaeggi et al., 2008; Taatgen, 2013, 2016).

Another interesting theoretical insight is offered by the concurrent presence of some near-transfer effects and absence of far-transfer effects, especially on measures of fluid intelligence. As seen, fluid intelligence and WM have been claimed to share similar mechanisms (e.g., shared capacity constraint; Jaeggi et al., 2008). Our results do not support this hypothesis. By contrast, our findings are in line with more recent evidence suggesting that WM and fluid intelligence differ from each other in terms of the underlying neural mechanisms (up-regulation and down-regulation of modularity; Lebedev, Nilsson, & Lövdén, 2018). The present meta-analysis, therefore, corroborates the hypothesis according to which WM and fluid intelligence are two non-isomorphic cognitive constructs supported by distinct neural mechanisms.

From a practical point of view, the most relevant implication is that

WM-training programs do not improve overall cognitive function in healthy older adults. Thus, they cannot be recommended as tools for slowing down cognitive decline or restoring overall cognitive ability. That said, less clear is the position to take with regard to the near-transfer effects induced by WM training. As pointed out earlier, increased memory skills may have a significant impact on older adults' quality of life, even without any far-transfer effect. However, two aspects of our results cast some doubt on the benefits of WM training for older adults' memory skills. First, the corrected (i.e., unbiased) near-transfer effect was relatively small (between $\overline{g} = 0.159$ and $\overline{g} = 0.246$). Second, the fact that the size of near-transfer effects is analogous in younger and older adults (about 0.150–0.250 SMD) suggests that the mechanisms underlying transfer do not depend on intrinsic features of WM plasticity. While training effects (i.e., criterion effects) appear to slightly decrease with age, possibly as a result of neural plasticity decline, no such pattern of results is observed in near-transfer measures. In our opinion, this state of affairs upholds the idea that WM training does not impact on WM capacity as a domain-general cognitive mechanism. Rather, such small effects can be easily accounted for by assuming that participants learn how to carry out a particular class of memory tasks rather than enhance their memory capacity (Gathercole et al., 2019; Melby-Lervåg et al., 2016; Sala & Gobet, 2019; Shipstead, Redick, & Engle, 2012). Put simply, practicing WM tasks may help to develop the ability to perform similar tasks (e.g., free recall, *n*-back, and span tasks) without having any impact on domain-general cognitive constructs such as short-term or long-term memory. If so, the consequent lack of transfer from the laboratory to real-life tasks would make WM training of little practical interest. Based on the available empirical evidence and due to the doubts about the nature of the observed near transfer and its small size, we conclude that WM training, to date, should not be recommended as a tool for improving or restoring older adults' memory skills. Therefore, the near-transfer effects following WM training programs may not be worth the effort and, possibly, should be abandoned in favor of other types of intervention, such as teaching mnemonics (e.g., Hertzog, Lövdén, Lindenberger, & Schmiedek, 2017; McCabe, Redick, & Engle, 2016; Verhaeghen, Marcoen, & Goossens, 1992). That being said, we think that further research is needed to clarify this issue.

5.2. Recommendations for future research

Given the evidence produced so far, further searching for the benefits of WM training on domain-general cognitive skills seems unproductive regardless of the particular population under examination. This appears particularly obvious if we consider that (a) no appreciable effect has been obtained in any of the far-transfer outcome measures and (b) no true heterogeneity suggesting differential effects has been found. Rather, the field should focus on clarifying the actual size and nature of the observed near-transfer effects. We present some suggestions for improving the methodological quality of WM training experiments with older adults.

Due to the similarities between trained tasks and some memory tests, the latter may not be adequate proxies for domain-general and transferable memory skills. To address this issue, future studies should include multivariate measures of short-term memory and WM. A set of several measures is necessary to investigate the effects of WM training on latent factors representing cognitive skills rather than single tasks (Noack et al., 2009; Schmiedek, Lövdén, & Lindenberger, 2010; Shipstead et al., 2012). Structural equation modeling (SEM) is the ideal approach for discriminating between task-related and factor-related effects. Assuming that WM training does improve WM as a core cognitive mechanism, and not only the ability to perform some memory tasks, we should expect the training-related improvements to occur through a latent factor exhibiting longitudinal strict measurement invariance. Simply put, since the baseline latent factor represents a particular domain-general memory skill (e.g., WM capacity), we should expect the post-test latent factor to measure the construct with the same metric, scale, and precision as at pre-test assessment. If this condition is not met, then it would be reasonable to conclude that WM training does not enhance domain-general memory skills. We realize that administering many cognitive tests and recruiting large samples—which are necessary to power a structural model—require significant financial and organizational resources. Nonetheless, we think that these characteristics, together with rigorous experimental designs, are necessary to produce robust findings that can contribute to our theoretical knowledge of the phenomenon and provide reliable advice for policy-makers.

Acknowledgements

We gratefully thank all the authors who provided unpublished data. We also thank Monica Melby-Lervåg for providing useful comments on an earlier draft. This work was supported by the Japan Society for the Promotion of Science [17F17313 granted to GS].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.intell.2019.101386.

References

Aksayli, N. D., Sala, G., & Gobet, F. (2019). The cognitive and academic benefits of Cogmed: A meta-analysis. *Educational Research Review, 29*, 229–243. https://doi.org/10.1016/j.edurev.2019.04.003.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA publications and communications board task force report. *American Psychologist, 73*, 3–25. https://doi.org/10.1037/amp0000191.

Archibald, L. M. D., & Gathercole, S. E. (2006). Short-term and working memory in specific language impairment. *International Journal of Language & Communication Disorders, 41*, 675–693. https://doi.org/10.1080/13682820500442602.

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*, 331–337. https://doi.org/10.3758/s13423-015-0967-4.

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*, 366–377. https://doi.org/10.3758/s13423-014-0699-x.

Baddeley, A. (1992). Working memory. *Science, 255*, 556–559. https://doi.org/10.1126/science.1736359.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*, 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2.

Begg, C. B., & Berlin, J. A. (1988). Publication bias: A problem in interpreting medical data. *Journal of the Royal Statistical Society, 151*, 419–463.

Boot, W. R., Simons, D. J., Stothart, C., & Stutts, C. (2013). The pervasive problem with placebos in psychology: Why active control groups are not sufficient to rule out placebo effects. *Perspectives on Psychological Science, 8*, 445–454. https://doi.org/10.1177/1745691613491271.

Carroll, J. B. (1993). *Human cognitive abilities.* Cambridge: University Press.

Carter, E. C., Schonbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science, 2*, 115–144. https://doi.org/10.1177/2515245919847196.

Chase, W. G., & Ericsson, K. A. (1982). Skill and working memory. In G. H. Bower (Ed.). *The psychology of learning and motivation* (pp. 1–58). New York: Academic.

Cheung, S. F., & Chan, D. K. (2014). Meta-analyzing dependent correlations: An SPSS macro and an R script. *Behavior Research Methods, 46*, 331–345. https://doi.org/10.3758/s13428-013-0386-2.

Conway, A. R. A., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review, 8*, 331–335. https://doi.org/10.3758/Bf03196169.

Dougherty, M. R., Hamovits, T., & Tidwell, J. W. (2016). Revaluating the effect of n-back training on transfer through the Bayesian lens: Support for the null. *Psychonomic Bulletin & Review, 23*, 306–316. https://doi.org/10.3758/s13423-015-0865-9.

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*, 276–284. https://doi.org/10.1111/j.0006-341X.2000.00455.x.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629–634. https://doi.org/10.1136/bmj.315.7109.629.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable

approach. *Journal of Experimental Psychology: General, 128*, 309–331. https://doi.org/10.1037/0096-3445.128.3.309.
Fisher, Z., Tipton, E., & Zhipeng, H. (2017). *Package "robumeta"*. Retrieved from https://cran.r-project.org/web/packages/robumeta/robumeta.pdf.
Gathercole, S. E., Dunning, D. L., Holmes, J., & Norris, D. (2019). Working memory training involves learning new skills. *Journal of Memory and Language, 105*, 19–42. https://doi.org/10.1016/j.jml.2018.10.003.
Gobet, F. (2016). *Understanding expertise: A multi-disciplinary approach*. London: Palgrave/Macmillan.
Gobet, F., & Simon, H. A. (1996). Templates in chess memory: A mechanism for recalling several boards. *Cognitive Psychology, 31*, 1–40. https://doi.org/10.1006/cogp.1996.0011.
Gray, J. R., Chabris, C. F., & Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience, 6*, 316–322. https://doi.org/10.1038/nn1014.
Green, C. S., Bavelier, D., Kramer, A. F., Vinogradov, S., Ansorge, U., Ball, K. K., ... Witt, C. M. (2019). Improving methodological standards in behavioral interventions for cognitive enhancement. *Journal of Cognitive Enhancement, 3*, 2–29. https://doi.org/10.1007/s41465-018-0115-y.
Guye, S., & von Bastian, C. C. (2017). Working memory training in older adults: Bayesian evidence supporting the absence of transfer. *Psychology and Aging, 32*, 732–746. https://doi.org/10.1037/pag0000206.
Halford, G. S., Cowan, N., & Andrews, G. (2007). Separating cognitive capacity from knowledge: A new hypothesis. *Trends in Cognitive Sciences, 11*, 236–242. https://doi.org/10.1016/j.tics.2007.04.001.
Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*, 39–65. https://doi.org/10.1002/jrsm.5.
Hempel, S., Miles, J. N. V., Booth, M. J., Wang, Z., Morton, S. C., & Shekelle, P. G. (2013). Risk of bias: A simulation study of power to detect study-level moderator effects in meta-analysis. *Systematic Reviews, 2*, 107.
Hertzog, C., Lövdén, M., Lindenberger, U., & Schmiedek, F. (2017). Age differences in coupling of intraindividual variability in mnemonic strategies and practice-related associative recall improvements. *Psychology and Aging, 32*, 557–571. https://doi.org/10.1037/pag0000177.
Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America, 105*, 6829–6833. https://doi.org/10.1073/pnas.0801268105.
Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General, 132*, 47–70. https://doi.org/10.1037/0096-3445.132.1.47.
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*, 66–71. https://doi.org/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 younger and older adults. *Psychological Science, 25*, 2027–2037. https://doi.org/10.1177/0956797614548725.
Kepes, S., Banks, G. C., & Oh, I.-S. (2014). Avoiding bias in publication bias research: The value of "null" findings. *Journal of Business and Psychology, 29*, 183–203. https://doi.org/10.1007/s10869-012-9279-0.
Kepes, S., & McDaniel, M. A. (2015). The validity of conscientiousness is overestimated in the prediction of job performance. *PLoS One, 10*, e0141468. https://doi.org/10.1371/journal.pone.0141468.
Klingberg, T. (2010). Training and plasticity of working memory. *Trends in Cognitive Sciences, 14*, 317–324. https://doi.org/10.1016/j.tics.2010.05.002.
Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlstrom, 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*, 177–186. https://doi.org/10.1097/00004583-200502000-00010.
Lampit, A., Hallock, H., & Valenzuela, M. (2014). Computerized cognitive training in cognitively healthy older adults: A systematic review and meta-analysis of effect modifiers. *PLoS Medicine, 11*, e1001756. https://doi.org/10.1371/journal.pmed.1001756.
Lebedev, A. V., Nilsson, J., & Lövdén, M. (2018). Working memory and reasoning benefit from different modes of large-scale brain dynamics in healthy older adults. *Journal of Cognitive Neuroscience, 30*, 1033–1046. https://doi.org/10.1162/jocn_a_01260.
McAvinue, L. P., Golemme, M., Castorina, M., Tatti, E., Pigni, F. M., Salomone, S., ... Robertson, I. H. (2013). An evaluation of a working memory training scheme in older adults. *Frontiers in Aging Neuroscience, 5*, 20. https://doi.org/10.3389/fnagi.2013.00020.
McCabe, J. A., Redick, T. S., & Engle, R. W. (2016). Brain-training pessimism, but applied-memory optimism. *Psychological Science in the Public Interest, 17*, 187–191. https://doi.org/10.1177/1529100616664716.
Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology, 49*, 270–291. https://doi.org/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*, 324–330. https://doi.org/10.3758/s13423-015-0862-z.
Melby-Lervåg, M., Redick, T. S., & 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*, 512–534. https://doi.org/10.1177/1745691616635612.
Mewborn, C. M., Lindbergh, C. A., & Miller, L. S. (2017). Cognitive interventions for cognitively healthy, mildly impaired, and mixed samples of older adults: A systematic review and meta-analysis of randomized-controlled trials. *Neuropsychology Review, 27*, 403–439. https://doi.org/10.1007/s11065-017-9350-8.
Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151*, 264–269. https://doi.org/10.7326/0003-4819-151-4-200908180-00135.
Moreau, D., Macnamara, B. N., & Hambrick, D. Z. (2018). Overstating the role of environmental factors in success: A cautionary note. *Current Directions in Psychological Science*. https://doi.org/10.1177/0963721418797300 Advanced online publication.
Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group design. *Organizational Research Methods, 11*, 364–386. https://doi.org/10.1177/1094428106291059.
Nguyen, L., Murphy, K., & Andrews, G. (2019). Immediate and long-term efficacy of executive functions cognitive training in older adults: A systematic review and meta-analysis. *Psychological Bulletin, 45*, 698–733. https://doi.org/10.1037/bul0000196.
Nilsson, J., Lebedev, A. V., Rydström, A., & Lövdén, M. (2017). Direct-current stimulation does little to improve the outcome of working memory training in older adults. *Psychological Science, 28*, 907–920. https://doi.org/10.1177/0956797617698139.
Noack, H., Lövdén, M., Schmiedek, F., & Lindenberger, U. (2009). Cognitive plasticity in adulthood and old age: Gauging the generality of cognitive intervention effects. *Restorative Neurology and Neuroscience, 27*, 435–453. https://doi.org/10.3233/RNN-2009-0496.
Passolunghi, M. C. (2006). Working memory and arithmetic learning disability. In T. P. Alloway, & S. E. Gathercole (Eds.). *Working memory and neurodevelopmental condition* (pp. 113–138). Hove, England: Psychology Press.
Peijnenborgh, J. C. A. W., Hurks, P. M., Aldenkamp, A. P., Vles, J. S. H., & Hendriksen, J. G. M. (2016). Efficacy of working memory training in children and adolescents with learning disabilities: A review study and meta-analysis. *Neuropsychological Rehabilitation, 26*, 645–672. https://doi.org/10.1080/09602011.2015.1026356.
Peng, P., Barnes, M., Wang, C., Wang, W., Li, S., Swanson, H. L., ... Tao, S. (2018). A meta-analysis on the relation between reading and working memory. *Psychological Bulletin, 144*, 48–76. https://doi.org/10.1037/bul0000124.
Peng, P., Namkung, J., Barnes, M., & Sun, C. Y. (2016). A meta-analysis of mathematics and working memory: Moderating effects of working memory domain, type of mathematics skill, and sample characteristics. *Journal of Educational Psychology, 108*, 455–473. https://doi.org/10.1037/edu0000079.
Pergher, V., Shalchy, M. A., Pahor, A., Van Hulle, M. M., Jaeggi, S. M., & Seitz, A. R. (2019). Divergent research methods limit understanding of working memory training. *Journal of Cognitive Enhancement*. https://doi.org/10.1007/s41465-019-00134-7 Advanced online publication.
Rebok, G. W., Ball, K., Guey, L. T., Jones, R. N., Kim, H.-Y., King, J. W., ... Willis, L. S. (2014). Ten-year effects of the advanced cognitive training for independent and vital elderly cognitive training trial on cognition and everyday functioning in older adults. *Journal of the American Geriatrics Society, 62*, 16–24. https://doi.org/10.1111/jgs.12607.
Redick, T. S., Calvo, A., Gay, E. G, & Engle, R. W. (2011). Working Memory Capacity and Go/No-Go Task Performance: Selective Effects of Updating, Maintenance, and Inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 308–324. https://doi.org/10.1037/a0022216.
Sala, G., Aksayli, N. D., Tatlidil, K. S., Tatsumi, T., Gondo, Y., & Gobet, F. (2019). Near and far transfer in cognitive training: A second-order meta-analysis. *Collabra: Psychology, 5*, 18. https://doi.org/10.1525/collabra.203.
Sala, G., & Gobet, F. (2017). Working memory training in typically developing children: A meta-analysis of the available evidence. *Developmental Psychology, 53*, 671–685. https://doi.org/10.1037/dev0000265.
Sala, G., & Gobet, F. (2019). Cognitive training does not enhance general cognition. *Trends in Cognitive Sciences, 23*, 9–20. https://doi.org/10.1016/j.tics.2018.10.004.
Salthouse, T. A. (2009). *Relations between age and cognitive functioning major issues in cognitive aging*. New York, NY: Oxford University Presshttps://doi.org/10.1093/acprof:oso/9780195372151.001.0001.
Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Newbury Park, CA: Sage.
Schmidt, F. L., & Oh, I.-S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology, 4*, 32–37. https://doi.org/10.1037/arc0000029.
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. https://doi.org/10.3389/fnagi.2010.00027.
Schwaighofer, M., Fischer, F., & Buhner, M. (2015). Does working memory training transfer? A meta-analysis including training conditions as moderators. *Educational Psychologist, 50*, 138–166. https://doi.org/10.1080/00461520.2015.1036274.
Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin, 138*, 628–654. https://doi.org/10.1037/a0027473.
Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. L. (2016). Do "brain-training" programs work? *Psychological Science in the Public Interest, 17*, 103–186. https://doi.org/10.1177/1529100616661983.
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*, 666–681. https://doi.org/10.1177/1745691614553988.
Soveri, A., Antfolk, J., Karlsson, L., Salo, B., & Laine, M. (2017). Working memory training

revisited: A multi-level meta-analysis of n-back training studies. *Psychonomic Bulletin & Review, 24*, 1077–1096. https://doi.org/10.3758/s13423-016-1217-0.

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science, 8*, 581–591. https://doi.org/10.1177/1948550617693062.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5*, 60–78. https://doi.org/10.1002/jrsm.1095.

Swanson, H. L. (2006). Working memory and reading disabilities: Both phonological and executive processing deficits are important. In T. P. Alloway, & S. E. Gathercole (Eds.). *Working memory and neurodevelopmental disorders* (pp. 59–88). Hove, England: Psychology Press.

Taatgen, N. A. (2013). The nature and transfer of cognitive skills. *Psychological Review, 120*, 439–471. https://doi.org/10.1037/a0033138.

Taatgen, N. A. (2016). Theoretical models of training and transfer effects. In T. Strobach, & J. Karbach (Eds.). *Cognitive training: An overview of features and applications* (pp. 19–29). New York, NY: Springer. https://doi.org/10.1007/978-3-319-42662-4_3.

Teixeira-Santos, A. C., Moreira, C. S., Magalhaes, R., Magalhaes, C., Pereira, D. R., Leite, J., ... Sampaio, A. (2019). Reviewing working memory training gains in healthy older adults: A meta-analytic review of transfer for cognitive outcomes. *Neuroscience and Biobehavioral Reviews*. https://doi.org/10.1016/j.neubiorev.2019.05.009 Advanced online publication.

Tetlow, A. M., & Edwards, J. D. (2017). Systematic literature review and meta-analysis of commercially available computerized cognitive training among older adults. *Journal of Cognitive Enhancement, 1*, 559–575. https://doi.org/10.1007/s41465-017-0051-2.

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114*, 104–132. https://doi.org/10.1037/0033-295X.114.1.104.

Verhaeghen, P., Marcoen, A., & Goossens, L. (1992). Improving memory performance in the aged through mnemonic training: A meta-analytic study. *Psychology and Aging, 7*, 242–251. https://doi.org/10.1037/0882-7974.7.2.242.

Viechtbauer, W. (2010). Conducting meta-analysis in R with the metafor package. *Journal of Statistical Software, 36*, 1–48. Retrieved from http://brieger.esalq.usp.br/CRAN/web/packages/metafor/vignettes/metafor.pdf.

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*, 112–125. https://doi.org/10.1002/jrsm.11.

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*, 190–212. https://doi.org/10.1037/neu0000227.