The cognitive and academic benefits of Cogmed: a meta-analysis

Article:

Aksayli, N. Deniz, Sala, Giovanni and Gobet, Fernand (2019) The cognitive and academic benefits of Cogmed: a meta-analysis. Educational Research Review, 27. pp. 229-243. ISSN 1747-938X

https://doi.org/10.1016/j.edurev.2019.04.003

Educational Research Review 27 (2019) 229–243

Contents lists available at ScienceDirect

Educational Research Review

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

Review

The cognitive and academic benefits of Cogmed: A meta-analysis

N. Deniz Aksayli[a], Giovanni Sala[b,*], Fernand Gobet[c]

[a] University of Nottingham, UK
[b] Osaka University, Japan
[c] University of Liverpool, UK

ARTICLE INFO

Keywords:
Cogmed
Meta-analysis
Transfer
Working memory training

ABSTRACT

Cogmed Working Memory Training (CWMT) is a commercial cognitive-training program designed to foster working-memory capacity. Enhanced working-memory capacity is then supposed to increase one's overall cognitive function and academic achievement. This meta-analysis investigates the effects of CWMT on cognitive and academic outcomes. The inclusion criteria were met by 50 studies (637 effect sizes).

Highly consistent near-zero effects were estimated in far-transfer measures of cognitive ability (e.g., attention and intelligence) and academic achievement (language ability and mathematics). By contrast, slightly heterogeneous small to medium effects were observed in memory tasks (i.e., near transfer). Moderator analysis showed that these effects were weaker for near-transfer measures not directly related to the trained tasks. These results highlight that, while near transfer occurs regularly, far transfer is rare or, possibly, inexistent. Transfer thus appears to be a function of the degree of overlap between trained tasks and outcome tasks.

Measures of cognitive ability, such as tests of reasoning and intelligence, are major predictors of academic and professional performance (Deary, Strand, Smith, & Fernandes, 2007; Detterman, 2014, 2016; Gobet, 2016; Rindermann & Neubauer, 2004; Schmidt, 2017). In recent years, the idea that experience can shape one's general cognitive skills, and, hence, positively impact people's lives, has brought forward the development of cognitive-training programs. Such programs have become a primary focus of interest in the field of cognitive psychology (Melby-Lervåg & Hulme, 2013; Simons et al., 2016).

The main idea behind cognitive-training programs is that core cognitive mechanisms can be boosted by engaging in cognitively demanding tasks/activities (for a review, see Strobach & Karbach, 2016). It is also theorized that the improvements specific to a particular cognitive domain impact on other non-trained cognitive and real-life skills (Jaeggi, Buschkuehl, Jonides, & Perrig, 2008; Jerrim, Macmillan, Micklewright, Sawtell, & Wiggins, 2016; Libertus et al., 2017; Rebok et al., 2014; Schellenberg, 2004). Thus, cognitive-training programs are believed to induce generalization of trained skills across a broad range of domains (i.e., far transfer) via enhancement of domain-general cognitive skills (Taatgen, 2016). This is the case of working-memory (WM) training.

WM is the ability to store and manipulate the information needed to perform cognitive tasks (Baddeley, 1992). WM capacity—that is, the maximum number of items WM can manipulate—has been associated with various cognitive functions, such as fluid intelligence and attention (Ackerman, Beier, & Boyle, 2005; Conway, Cowan, & Bunting, 2001; Engle, 2018; Kane, Hambrick, & Conway, 2005; Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002). WM capacity appears to be a significant predictor of academic achievement as well (Peng et al., 2018; Peng, Namkung, Barnes, & Sun, 2016). Moreover, low WM capacity is a correlate of learning disabilities such as attention-deficit hyperactivity disorder (ADHD; Westerberg, Hirvikoski, Forssberg, & Klingberg, 2004), language

[*] Corresponding author. Graduate School of Human Sciences, Yamada Oka 1-2 Suita, Osaka University, 565-0871, Japan.
E-mail address: sala@hus.osaka-u.ac.jp (G. Sala).

https://doi.org/10.1016/j.edurev.2019.04.003
Received 28 July 2018; Received in revised form 5 February 2019; Accepted 23 April 2019
Available online 08 May 2019

impairment (Archibald & Gathercole, 2006), and poor academic achievement (Passolunghi, 2006; Swanson, 2006). Given these premises, it is natural to assume that if WM capacity could be enhanced by training, the benefits would spread across many other skills positively related to WM (Jaeggi et al., 2008). This may be even truer for those people whose WM capacity is low or has been impaired such as children with learning disabilities (e.g., Klingberg, Forssberg, & Westerberg, 2002) and brain-injured patients (for a review see Weicker, Villringer, & Thöne-Otto, 2016).

The hypothesis underlying this claim is that WM and fluid intelligence have a shared capacity constraint (Halford, Cowan, & Andrews, 2007). The performance on fluid intelligence tasks (e.g., Raven's progressive matrices) is constrained by the amount of information (i.e., number of items) that can be manipulated by WM. If WM capacity is increased, then an improvement in such tasks is expected to occur (Jaeggi et al., 2008). In turn, improving fluid intelligence would benefit other cognitive and academic skills. Another complementary explanation of the presumed broad generalization of skills following WM training refers to the role played by attentional processes in both fluid intelligence and working memory tasks (Engle, 2018; Gray, Chabris, & Braver, 2003). Engaging in cognitively demanding activities such as WM training task may enhance attentional control, which is positively related with performance in most cognitive and academic tests. Therefore, like with the other cognitive-training programs, the essential assumption underlying these hypotheses is that WM training fosters domain-general mechanisms such as WM capacity and attentional control, which in turn enhances other cognitive and academic skills.

1. Cogmed Working Memory Training

Following this idea, some commercial cognitive-training computerized programs have been designed in the last two decades to boost WM capacity, overall cognitive ability, and everyday functioning. The most well-known, studied, and influential of such programs is Pearson's Cogmed Working Memory Training (hereafter CWMT; www.cogmed.com). Simons et al. (2016) classify CWMT as one of the five commercial cognitive-training programs whose effectiveness had been assessed in several publications (see also SharpBrains, 2015). The studies investigating the effects of CWMT are undoubtedly the most numerous and best designed (e.g., often including active controls) in this category (Simons et al., 2016, pp. 143–148).

CWMT is usually administered by school personnel or clinical practitioners who have been trained by Cogmed coaches. All the three types of CWMT programs—Cogmed JM for preschoolers, Cogmed RM for older children, and Cogmed QM for adults—consist of 25, 30, and 45-min sessions over a five-week period. Trainees perform the training either in a school or rehabilitation environment, or at home under remote supervision (Simons et al., 2016). The training regimens include gamified verbal and visuo-spatial WM tasks that require trainees to recall increasingly longer sequences of information as their performance improves with practice (for more details, see Shipstead, Hicks, & Engle, 2012a; Shipstead, Redick, & Engle, 2012b; Simons et al., 2016).

The preschool training, Cogmed JM, involves training tasks that are linked together with a theme-park design. Preschoolers are required to direct their attention towards a sequence of items (i.e. array of bright colored fur ball creatures), hold the sequence in their WM, and then select the items in their original order using a mouse or a touch pad. The duration of the WM tasks gets adjusted based on the trainee's performance. Correct responses get reinforced with positive visual stimuli (e.g., smiles) and the duration of the intervals between stimuli and recall increases, whereas incorrect responses elicit negative visual stimuli (e.g., frowning). The school-aged children training, Cogmed RM, involves training tasks presented on a space-themed interface design. Similar to Cogmed JM, trainees are required to recall a sequence of items from memory. However, the tasks in Cogmed RM have more targets and longer sequences, thus they are relatively more difficult than Cogmed JM tasks. In addition, trainees' scores are presented on the screen so that they can challenge themselves and try to outperform their previous score. As an incentive and reward for task completion, a racing game called “RoboRacing”, which involves collecting coins and racing against the clock, is presented at the end of each daily training. The adult training, Cogmed QM, involves similar training exercises as Cogmed RM. However, in this version, trainees may need to concentrate more because the interface is less visually appealing and the emphasis on surpassing prior performance is less apparent. For more details, see Roche and Johnson (2014).

Notably, the software has been claimed to increase performance in academic, social, and professional settings (www.cogmed.com/how-is-cogmed-different). Cogmed avers that CWMT leads to improvements in attention, reading, mathematics, cognitive control, and cognitive functioning in daily life (Pearson, 2016). Nonetheless, Cogmed also acknowledges that further and more compelling evidence is needed, especially with regard to the presumed academic benefits of CWMT.

As in most cognitive-training programs, the findings regarding CWMT have been mixed, which has kept researchers from reaching a definite conclusion on the topic. While some scientists have expressed optimism due to promising results (e.g., Shinaver, Entwistle, & Söderqvist, 2014), others have highlighted the overall insufficient quality of the experimental design of the studies investigating the effects of CWMT (e.g., Simons et al., 2016). The lack of active controls (or any controls), non-random allocation of the participants to the groups, and small sample sizes are some of the major flaws that may bias the results of CWMT studies and introduce spurious variability in the pool of data.

2. The present study

This paper evaluates the impact of CWMT on people's cognitive and academic skills via meta-analysis. We focus on two primary goals. First, we evaluate the differential impact of CWMT on performance in cognitive tasks as a function of the type of transfer. The field has customarily distinguished between far-transfer and near-transfer effects (Barnett & Ceci, 2002). While the latter concern the performance on memory tasks (as proxies for WM capacity), the former deal with cognitive and academic tasks (e.g., fluid intelligence, attention, language, and mathematics). Furthermore, several memory tasks employed as outcome measures closely resemble the tasks included in CWMT. The training is thus expected to exert a stronger effect on such tasks rather than those memory tasks that are not directly related to the trained tasks.

Second, we aim to quantify and explain the amount of variability in the findings in this literature. We employ moderator analysis to investigate the potential sources of within- and between-study heterogeneity. This analysis addresses a fundamental point: statistically accounting for the degree of true heterogeneity is the only reliable way to make some sense of the mixed results the field has produced so far.

To the best of our knowledge, only two meta-analyses have been carried out specifically on CWMT so far. Spencer-Smith and Klingberg's (2015) meta-analysis is somewhat limited in scope (it deals solely with subjective measures of inattention in daily life). Nutley & Ralph's (n.d.) meta-analysis is relatively outdated (only 16 studies are included, and the most recent ones are from 2012). Both these meta-analytic syntheses have reported results supporting the effectiveness of CWMT. Other meta-analyses have included some CWMT interventions within the broader context of WM training, showing less optimistic results (e.g., Melby-Lervåg, Redick, & Hulme, 2016; Sala & Gobet, 2017a). However, it is worth mentioning that the number of CWMT interventions included in these latter meta-analyses is quite limited (19 in Melby-Lervåg et al., 2016; four in Sala & Gobet, 2017a) and no conclusive response about CWMT effectiveness can be drawn from these syntheses. In the last few years, the number of eligible studies has more than doubled, and no previous meta-analysis has in fact been conclusive regarding the actual impact of CWMT on cognitive ability and academic skills. Given the large amount of experimental evidence collected so far, the prominent position of CWMT in the landscape of commercial cognitive-training programs, the importance of theoretical and potential practical (clinical and educational) implications, and the contradictory claims expressed by different researchers in the field, we think that an up-to-date meta-analytic synthesis implementing a sound modeling design is required.

3. Method

3.1. Literature search

A systematic search strategy was employed to find the relevant studies (PRISMA statement; Moher, Liberati, Tetzlaff, & Altman, 2009). The following Boolean string was used: ("Cogmed" OR "working memory training" OR "WM training" OR "cognitive training"). We searched through MEDLINE, PsycINFO, Science Direct, and ProQuest Dissertation & Theses databases to identify all the potentially relevant studies. We retrieved 2761 records. Also, earlier meta-analytic and narrative reviews were examined (e.g., Melby-Lervåg et al., 2016; Nutley & Ralph, n.d.; Pearson, 2016; Simons et al., 2016; Villemonteix, 2018).

3.2. Inclusion criteria

The studies were included according to the following five criteria:

1. The study included at least one group trained *solely* on CWMT and at least one control group not engaged in adaptive CWMT or any other adaptive WM-training program. This criterion was fundamental to isolate the variable of interest, that is, the impact of CWMT on performance in cognitive/academic tasks;
2. At least one cognitive/academic task was administered. Self-reported measures and parent/teacher rating questionnaires were excluded.[1] Also, when the control group was involved in activities closely related to one of the outcome measures (e.g., controls involved in a math course), the relevant effect sizes were excluded (e.g., tests of mathematical achievement);
3. The study included both pre- and post-test assessments;
4. The study reported new data (i.e., it did not report only duplicate results from previous studies);
5. The data reported in the study (or provided by the author) were sufficient to compute an effect size.

We searched for eligible published and unpublished articles through December 31st, 2017. We sent emails ($n = 11$) to researchers in the field asking for the necessary data to calculate the effect sizes. We received three positive replies. In total, we found 50 studies, conducted from 2005 to 2017, that met all the inclusion criteria. These studies included 637 effect sizes and a total of 3059 participants. The entire procedure is described in Fig. 1. The Supplemental materials available online contain the details of all the included studies and a list of the excluded studies.

3.3. Meta-analytic models

Each effect size was considered either *near-transfer* or *far-transfer*. The near-transfer effect sizes consisted of memory tasks referring to the Gsm construct as defined by the Cattell-Horn-Carroll model (CHC model; McGrew, 2009). Far-transfer effect sizes referred to all the other cognitive measures (for the details, see 3.4. Moderators section and the Supplemental materials available online). Two authors coded each effect size independently and reached 100% agreement.

[1] Despite being good proxies for daily life cognitive ability, subjective measures are sensitive to expectancy effects in true experiments. For more details, see Cortese et al. (2015), Rapport, Orban, Kofler, and Friedman (2013), Simons et al. (2016), and Sonuga-Barke et al. (2013).

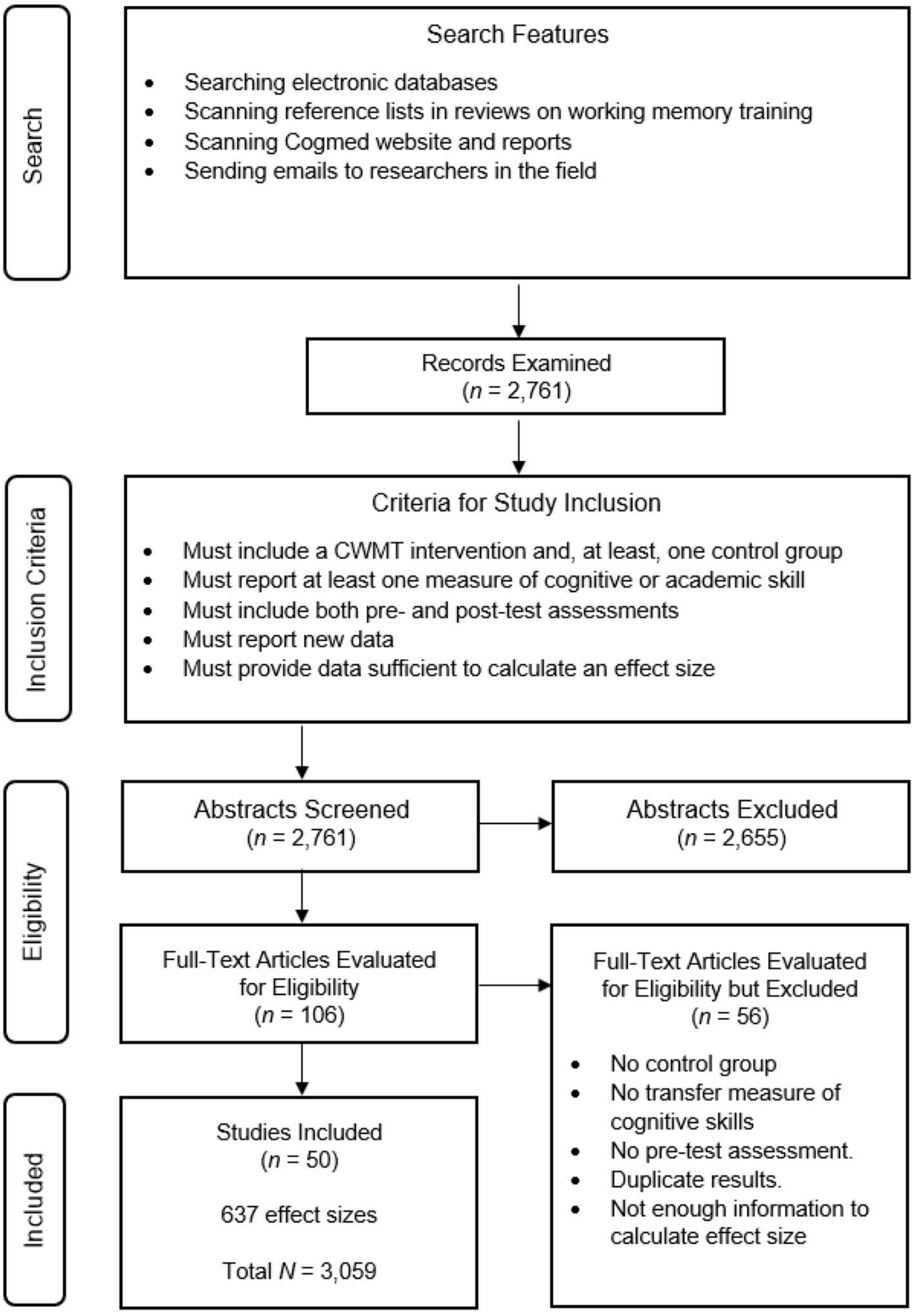
Search
Search Features
Searching electronic databases
Scanning reference lists in reviews on working memory training
Scanning Cogmed website and reports
Sending emails to researchers in the field
Records Examined (n = 2,761)
Inclusion Criteria
Criteria for Study Inclusion
Must include a CWMT intervention and, at least, one control group
Must report at least one measure of cognitive or academic skill
Must include both pre- and post-test assessments
Must report new data
Must provide data sufficient to calculate an effect size
Eligibility
Abstracts Screened (n = 2,761)
Abstracts Excluded (n = 2,655)
Full-Text Articles Evaluated for Eligibility (n = 106)
Full-Text Articles Evaluated for Eligibility but Excluded (n = 56)
No control group
No transfer measure of cognitive skills
No pre-test assessment.
Duplicate results.
Not enough information to calculate effect size
Included
Studies Included (n = 50)
637 effect sizes
Total N = 3,059

Fig. 1. Flow diagram of the search strategy.

3.4. Moderators

We chose seven potential moderators:

1. Allocation (dichotomous variable): Whether the participants were randomly assigned to the experimental and control groups;
2. Control group (active or non-active; dichotomous variable): Whether the CWMT group was compared to another cognitively demanding activity (e.g., non-adaptive training); no-contact groups and business-as-usual groups were considered as “non-active.” Also, in one study (Hadwin & Richards, 2016) the control group was involved in non-cognitive tasks (cognitive behavioral therapy). This control group was labeled as “non-active” as well;
3. Baseline difference (continuous variable): The standardized mean difference corrected for upward bias (i.e., Hedges's *g*; see 3.5. Effect Size Calculation) between the experimental and control groups at pre-test assessment. This moderator was included to

check whether the part of the observed heterogeneity was due to regression to the mean;

4. Age (categorical variable): Whether the participants were *children* (16-year-old or younger), *adults* (17–55-year-old), or *older adults* (older than 55)[2];
5. Population (dichotomous variable): Whether the participants were typical subjects not suffering from any clinical conditions (e.g., ADHD) or intellectual disabilities;
6. Measure (categorical variable): This moderator, which was added only in the far-transfer models, included (a) measures of cognitive skills such as fluid intelligence (Gf in the CHC nomenclature) and attentional skills (Gs/Gt); (b) measures of academic skills such as language ability (Language) and mathematical ability (Math); and (c) full-scale IQ (i.e., batteries including tests of verbal and non-verbal intelligence, and sometimes tests of Gs/Gt). Those effect sizes that did not fall into any of the above categories were labeled as (d) "miscellaneous." The first 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;
7. Criterion (categorical variable): Whether the task resembled one of the training tasks in CWMT (*very near transfer*) or was a different memory task (*lesser near transfer*). This moderator was added only in the near-transfer models. The first two authors coded each effect size for moderator variables independently. The inter-rater agreement was 98%. The two authors resolved every discrepancy by discussion.

3.5. Effect Size Calculation

The effect sizes were calculated for each eligible task reported in the primary studies. Those effect sizes that were redundant (e.g., sum of digit span forward and backward when the individual indexes were reported) were excluded.

The effect size used was Hedges's *g*. The formula for the effect size was:

$$g = \frac{(M_{e\_post} - M_{e\_pre}) - (M_{c\_post} - M_{c\_pre})}{SD_{pooled_{pre}}} \times \left(1 - \frac{3}{(4 \times N) - 9}\right) \tag{1}$$

where M_{e_post} and M_{e_pre} are the mean performance of the experimental group at post-test and pre-test, respectively, M_{c_post} and M_{c_pre} are the mean performance of the control group at post-test and pre-test, respectively, SD_{pooled_pre} is the pooled pre-test SDs in the experimental group and the control group, and N is the total sample size.

The formula for the corresponding sampling error variance of the effect size was[3]:

$$Var_g = \left(\frac{N}{N_e \times N_c} + \frac{d^2}{2 \times N}\right) \times \left(1 - \frac{3}{(4 \times N) - 9}\right)^2 \tag{2}$$

where d is the standardized mean difference (i.e., the first factor of Equation (1)), N_e the size of the experimental group, and N_c the size of the control group (Hedges & Olkin, 1985; Schmidt & Hunter, 2015, pp. 292–293).

3.6. Modeling approach

We employed *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; Tanner-Smith & Tipton, 2014; Tanner-Smith, Tipton, & Polanin, 2016). RVE models nested effect sizes (i.e., effect sizes extracted from the same study) and calculates robust standard errors. RVE also estimates the within-cluster-variance (ω^2) and between-cluster-variance components (τ^2) expressing the amount of true heterogeneity in the dataset. We thus grouped all the effect sizes extracted from one study into the same cluster. The Robumeta software R package (Fisher, Tipton, & Zhipeng, 2017) was used to run the analyses.

3.7. Sensitivity analysis

To test the robustness of the results, we performed Viechtbauer and Cheung's (2010) influential case analysis (run with Metafor R package; Viechtbauer, 2010). This analysis evaluated whether some effect sizes were outliers or exerted an unusually strong influence on the overall effect sizes.[4] The meta-analytic models were thus run both with and without influential effect sizes.

Once the influential effect sizes were removed, we used Cheung and Chan's (2014) weighted-samplewise correction to merge the effect sizes extracted from the same paper. (For more details on the procedure, see the Supplemental material available online.) We then ran several

[2] The type of CWMT (JM, RM, and QM) was not added as a moderator because it was confounded with age.

[3] It is worth noting that the most accurate formula for the calculation of sampling error variance in repeated measures designs with control groups requires pre-post-test correlations (Schmidt & Hunter, 2015, pp. 343–355). Such information is rarely provided in the included primary studies (only Honoré & Noël, 2017 report pre-post-test correlations). That said, we think that the formula we used is an acceptable approximation. In the supplemental materials, we report the R codes to reproduce the results with the Schmidt and Hunter's (2015) formula assuming a realistic pre-post-test correlation ($r = 0.650$). The only appreciable difference is a further reduction of the amount of true heterogeneity.

[4] A few very-near-transfer effect sizes were excessively large (e.g., $g > 2$). However, the influential case analysis did not detect them because the relevant sampling variances were too high. These effect sizes were excluded. For the details, see 4. Results section.

publication bias analyses.[5] Running multiple publication bias analyses is recommended to test the robustness of the naïve (i.e., uncorrected) estimates (Kepes & McDaniel, 2015). First, we visually inspected the funnel plots for possible asymmetries. Second, we used the trim-and-fill analysis (Duval & Tweedie, 2000). This method estimates whether some smaller-than-average effects have been suppressed and calculates a corrected overall effect size based on the asymmetry observed in the funnel plots. We used all the three estimators (*L0*, *R0*, and *Q0*) described in Duval and Tweedie (2000; run with Metafor R package). The three estimators differ from each other only regarding the type of non-parametric test they implement. Using three different estimators is recommended in order to increase the reliability of the corrected overall effect sizes. Finally, since trim-and-fill analysis 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 preferred over the PET estimator (Stanley, 2017; Stanley & Doucouliagos, 2014).

3.8. Follow-up effects

A subsample of the studies reported both immediate post-test effects and follow-up effects. Two studies (Foy & Mann, 2014; Roberts et al., 2016) reported only follow-up effects. The effect sizes were calculated by replacing the numerator in formula (1) with the difference between the follow-up mean and the pre-test mean in the two groups. The analyses described above were run for follow-up effects as well.

Furthermore, we ran additional analyses to test the robustness of the effects from post-test to follow-up. We included only those studies that tested the participants at both post-test and follow-up. These analyses are reported in the Supplemental materials available online (Tables S1–S4).

4. Results

4.1. Far transfer

4.1.1. Immediate post-test

The RVE model included all the effect sizes related to far-transfer measures, that is, those measures that shared no overlap with the training tasks. The overall effect size was $\bar{g} = 0.048$, $SE = 0.031$, 95% CI [-0.017; 0.113], $m = 39$, $k = 194$, $df = 16.62$, $p = .135$, $\omega^2 = 0.000$, $\tau^2 = 0.006$. We ran a meta-regression model including all the moderators. The only significant moderator was Baseline ($b = -0.345$, $SE = 0.082$, $p = .001$, $\omega^2 = 0.000$, $\tau^2 = 0.000$). This does not mean that the small positive effect found ($\bar{g} = 0.048$) is attributable to regression to the mean. In fact, the overall effect size at baseline was near-zero ($\bar{g} = 0.011$). Rather, only the low between-study heterogeneity observed ($\tau^2 = 0.006$) was affected by the differences at baseline.

We found one influential case ($g = 0.195$, $ID = 59$; see Supplemental materials available online). After excluding this effect, the overall effect size was $\bar{g} = 0.044$, $SE = 0.031$, 95% CI [-0.022; 0.111], $m = 38$, $k = 193$, $df = 15.92$, $p = .175$, $\omega^2 = 0.000$, $\tau^2 = 0.006$. Again, Baseline was the only significant moderator ($b = -0.345$, $SE = 0.082$, $p = .001$, $\omega^2 = 0.000$, $\tau^2 = 0.000$). These results showed that the overall far-transfer effect was not significantly different from zero and that the observed true heterogeneity was only due to a statistical artifact (i.e., regression to the mean). Since all the observed true heterogeneity was accounted for, no variance was left to be explained and thus no other potential moderator could have affected the outcomes.

With regard to publication bias, the funnel plot looked slightly asymmetrical (a few extreme effects were observed on the right of the mean but not on the left; Fig. 2).

The trim-and-fill estimates were $\bar{g} = 0.028$ ($SE = 0.033$, $p = .398$), $\bar{g} = 0.033$ ($SE = 0.033$, $p = .318$), and $\bar{g} = 0.028$ ($SE = 0.033$, $p = .398$) with the *L0*, *R0*, and *Q0* estimators, respectively. The PET estimate was $\bar{g} = -0.156$ ($SE = 0.103$, $p = .137$), and the PEESE estimate was $\bar{g} = -0.045$ ($SE = 0.048$, $p = .350$). Thus, while the PET estimator clearly provided an overcorrected estimate, the other estimates confirmed that the effect was not significantly different from zero.

4.1.2. Follow-up

The RVE model included all the effect sizes related to far-transfer measures at follow-up. The overall effect size was $\bar{g} = 0.051$, $SE = 0.031$, 95% CI [-0.018; 0.119], $m = 22$, $k = 91$, $df = 11.40$, $p = .131$, $\omega^2 = 0.001$, $\tau^2 = 0.000$. We ran a meta-regression model including all the moderators. The only significant moderator was Baseline ($b = -0.364$, $SE = 0.101$, $p = .006$, $\omega^2 = 0.000$, $\tau^2 = 0.000$). No influential case was found. Simply put, no difference with the follow-up assessment was observed (no appreciable effect or true heterogeneity).

The funnel plot looked approximatively symmetrical (Fig. 3).

The trim-and-fill estimates were $\bar{g} = 0.045$ ($SE = 0.033$, $p = .175$) with the *L0* and *Q0* estimators, and $\bar{g} = 0.049$ ($SE = 0.033$,

[5] Excluding influential cases usually decreases the degree of true heterogeneity. Since high degrees of true heterogeneity can sometimes bias publication-bias-corrected estimates (Schmidt & Hunter, 2015; Stanley, 2017), we ran the publication-bias analyses without the detected influential cases.

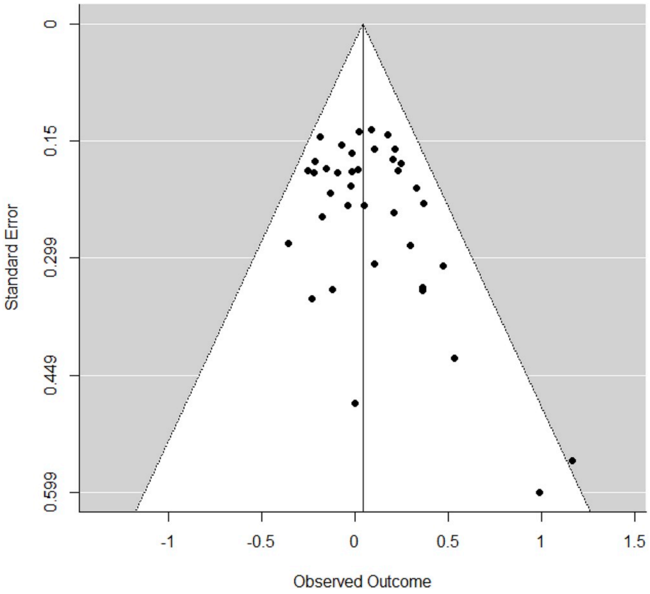

Fig. 2. Funnel plot of observed outcomes (*g*s) and standard errors of far-transfer measures.

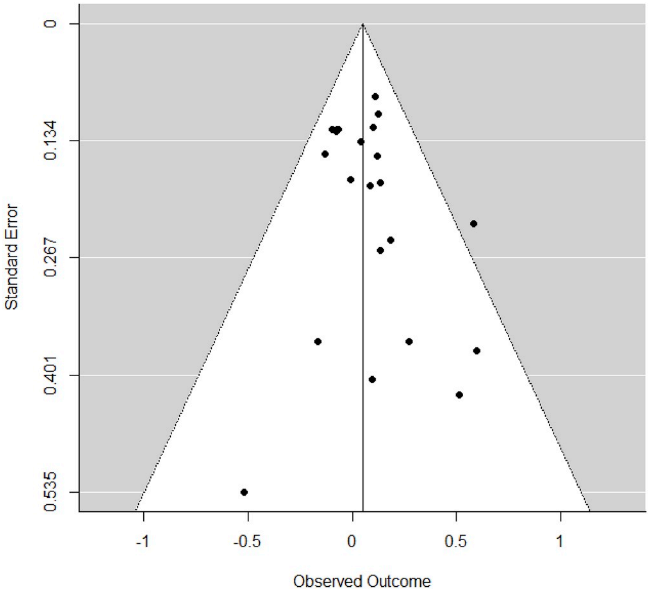

Fig. 3. Funnel plot of observed outcomes (*g*s) and standard errors of far-transfer measures at follow-up.

p = .138) with the *R0* estimator. The PET estimate was $\bar{g}\ = -0.023$ (SE = 0.074, p = .755), and the PEESE estimate was $\bar{g}\ = 0.027$ (SE = 0.040, p = .509). Thus, publication-bias analysis confirmed that follow-up overall effect was not significantly different from zero (all the estimates).

4.2. Near transfer

4.2.1. Immediate post-test

The RVE model included all the effect sizes related to near-transfer measures. The overall effect size was $\bar{g}\ = 0.444$, SE = 0.052, 95% CI [0.337; 0.551], m = 44, k = 247, df = 24.46, p < .001, ω^2 = 0.046, τ^2 = 0.071. We ran a meta-regression model including all the moderators. The significant moderators were Baseline (b = −0.487, SE = 0.137, p = .002) and Criterion (b = 0.330, SE = 0.054, p < .001). These two moderators explained about half of the observed variance (ω^2 = 0.024, τ^2 = 0.037).

We found two influential cases (g = 2.343, ID = 383 and g = 2.265, ID = 195). Another effect size was excluded because it was considered as an outlier (g = 2.152, ID = 552). After excluding these effects, the overall effect size was $\bar{g}\ = 0.427$, SE = 0.047, 95%

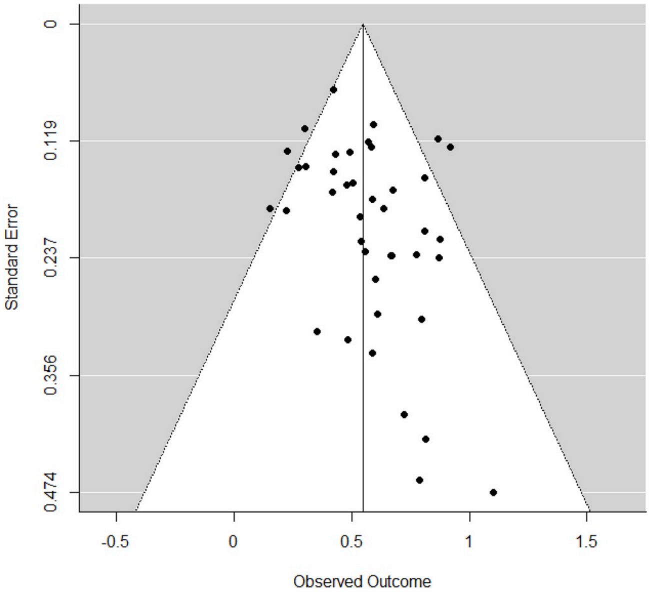

Fig. 4. Funnel plot of observed outcomes (*g*s) and standard errors of very-near-transfer measures.

CI [0.330; 0.524], $m = 43$, $k = 244$, $df = 24.21$, $p < .001$, $\omega^2 = 0.038$, $\tau^2 = 0.058$. Again, Baseline ($b = -0.450$, $SE = 0.106$, $p < .001$) and Criterion ($b = 0.328$, $SE = 0.055$, $p < .001$) were the only two significant moderators ($\omega^2 = 0.014$, $\tau^2 = 0.030$).

These analyses thus showed that the overall near-transfer effect was medium ($\bar{g} = 0.444$). This effect was robust to the exclusion of the influential cases. Finally, the effect was consistent. In fact, most of the observed true heterogeneity was accounted for by a few influential/extreme effects, the Baseline moderator (i.e., regression to the mean), and the degree of overlap between the trained task and the transfer task (moderator Criterion). The residual heterogeneity was low ($\omega^2 = 0.014$, $\tau^2 = 0.030$).

4.2.1.1. Analysis of criterion moderator. The degree of overlap between the training tasks of CWMT and the memory tasks was a significant moderator. We thus ran separate analyses for very-near-transfer measures and lesser-near-transfer measures. First, we ran an RVE model including all the effect sizes related to very-near-transfer measures. The overall effect size was $\bar{g} = 0.566$, $SE = 0.046$, 95% CI [0.472; 0.660], $m = 42$, $k = 154$, $df = 23.81$, $p < .001$, $\omega^2 = 0.061$, $\tau^2 = 0.040$. We ran a meta-regression model including all the moderators. No moderator was significant.

We found one influential case ($g = 2.265$, ID = 195). Another effect size was excluded because it was considered as an outlier ($g = 2.152$, ID = 552). After excluding these effects, the overall effect size was $\bar{g} = 0.550$, $SE = 0.042$, 95% CI [0.463; 0.637], $m = 41$, $k = 152$, $df = 23.57$, $p < .001$, $\omega^2 = 0.048$, $\tau^2 = 0.033$. Baseline ($b = -0.391$, $SE = 0.148$, $p = .015$) was the only significant moderator and the residual true heterogeneity was low ($\omega^2 = 0.029$, $\tau^2 = 0.033$).

These results showed that the participants' performance on those tasks that closely resembled the CWMT training tasks was greater than the overall near-transfer effect ($\bar{g} = 0.566$ and $\bar{g} = 0.444$, respectively). The model was homogeneous, especially when influential cases and outliers were removed, and baseline difference were controlled for.

The funnel plot looked slightly asymmetrical (Fig. 4).

The trim-and-fill estimates were $\bar{g} = 0.455$ ($SE = 0.039$, $p < .001$), $\bar{g} = 0.547$ ($SE = 0.036$, $p < .001$), and $\bar{g} = 0.433$ ($SE = 0.038$, $p < .001$) with the *L0*, *R0*, and *Q0* estimators, respectively. The PET estimate was $\bar{g} = 0.401$ ($SE = 0.075$, $p < .001$), and the PEESE estimate was $\bar{g} = 0.476$ ($SE = 0.042$, $p < .001$). In this case, the PET overcorrected because it was associated with a significant *p*-value (one-tailed $p < .100$). Thus, the PEESE estimator was more reliable. The publication-bias analysis thus suggested that the true overall effect in the very-near-transfer measures was slightly smaller than the uncorrected estimates (approximatively $\bar{g} = 0.450$ and $\bar{g} = 0.550$, respectively). Nonetheless, the effect was robust and highly significant.

Second, we ran an RVE model including all the effect sizes related to lesser-near-transfer measures. The overall effect size was $\bar{g} = 0.246$, $SE = 0.069$, 95% CI [0.102; 0.391], $m = 33$, $k = 93$, $df = 19.49$, $p = .002$, $\omega^2 = 0.000$, $\tau^2 = 0.091$. We ran a meta-regression model including all the moderators. Baseline was a significant moderator ($b = -0.598$, $SE = 0.169$, $p = .006$). Also, Age was a significant moderator, with the effect of the training significantly higher in the children than in the adults and older adults ($b = 0.318$, $SE = 0.117$, $p = .018$). These moderators explained nearly all the observed true heterogeneity ($\omega^2 = 0.000$, $\tau^2 = 0.018$). The overall effect size in the sample of children was $\bar{g} = 0.457$, $SE = 0.082$, $p < .001$. However, the higher effect size in children is probably due to a large extent to regression to the mean. In fact, the overall effect size at baseline was significantly negative ($\bar{g} = -0.198$, $SE = 0.054$, $p = .004$). In other words, this effect was a statistical artifact due to a certain amount of collinearity between Baseline and Age moderators.

We found one influential case ($g = 2.343$, $ID = 383$). After excluding this effect size, the overall effect size was $\bar{g} = 0.231$, $SE = 0.065$, 95% CI [0.094; 0.368], $m = 33$, $k = 92$, $df = 19.06$, $p = .002$, $\omega^2 = 0.000$, $\tau^2 = 0.066$. Baseline ($b = -0.506$,

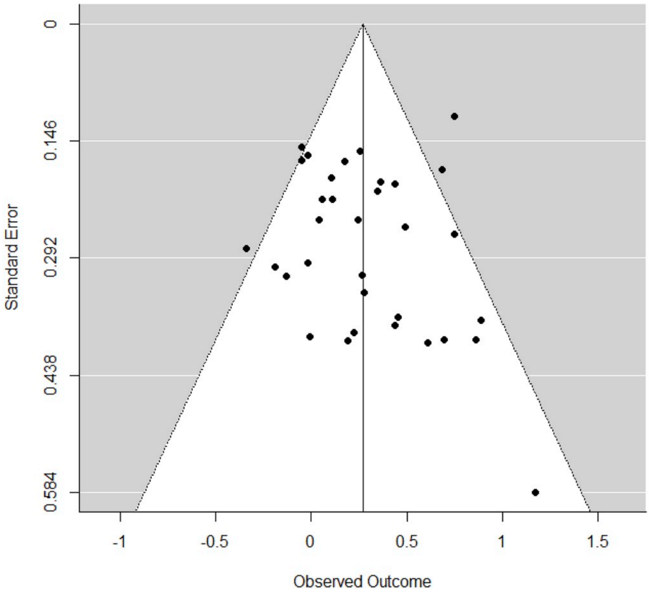

Fig. 5. Funnel plot of observed outcomes (*g*s) and standard errors of lesser-near-transfer measures.

$SE = 0.127$, $p = .004$) and Age (children, $b = 0.357$, $SE = 0.114$, $p = .009$) were significant moderators (but again, see the caveat above) and explained nearly all the observed true heterogeneity ($\omega^2 = 0.000$ and $\tau^2 = 0.001$).

The funnel plot looked symmetrical (Fig. 5).

The trim-and-fill estimates were $\bar{g} = 0.261$ ($SE = 0.058$, $p < .001$), $\bar{g} = 0.249$ ($SE = 0.058$, $p < .001$), and $\bar{g} = 0.261$ ($SE = 0.058$, $p < .001$) with the *L0*, *R0*, and *Q0* estimators, respectively. The PET estimate was $\bar{g} = 0.287$ ($SE = 0.160$, $p = .082$), and the PEESE estimate was $\bar{g} = 0.247$ ($SE = 0.088$, $p = .008$). Overall, no publication bias was observed because pretty much all the publication-biased corrected estimates were similar or slightly greater than the uncorrected overall effect size.

The analyses thus showed a very clear pattern of results. While the very-near-transfer overall effect was medium (about $\bar{g} = 0.450$ at least), the lesser-near-transfer effect was significantly smaller (about $\bar{g} = 0.250$ at most). This pattern of results was in line with the hypothesis according to which transfer is a function of the extent to which the trained task and the target task overlap (i.e., share common features).

4.2.2. *Follow-up*

The RVE model included all the effect sizes related to near-transfer measures at follow-up. The overall effect size was $\bar{g} = 0.365$, $SE = 0.057$, 95% CI [0.242; 0.487], $m = 23$, $k = 105$, $df = 13.62$, $p < .001$, $\omega^2 = 0.067$, $\tau^2 = 0.038$. We ran a meta-regression model including all the moderators. Baseline and Criterion were significant moderators ($b = -0.398$, $SE = 0.181$, $p < .050$ and $b = 0.359$, $SE = 0.085$, $p = .001$). Also, Age was a significant moderator, with the children performing worse than the adults and older adults ($b = -0.298$, $SE = 0.099$, $p = .029$). These moderators explained most of the within-cluster heterogeneity ($\omega^2 = 0.022$, $\tau^2 = 0.035$).

Two influential cases were found ($g = 1.998$, $ID = 99$ and $g = 1.647$, $ID = 87$). Another effect was excluded because it was a blatant outlier ($g = 3.062$, $ID = 553$). After excluding these effects, the overall effect size was $\bar{g} = 0.326$, $SE = 0.037$, 95% CI [0.247; 0.406], $m = 22$, $k = 102$, $df = 13.87$, $p < .001$, $\omega^2 = 0.065$, $\tau^2 = 0.000$. Baseline and Criterion were significant moderators ($b = -0.451$, $SE = 0.133$, $p = .006$ and $b = 0.322$, $SE = 0.060$, $p < .001$). The children underperformed compared to the adults and older adults ($b = -0.284$, $SE = 0.072$, $p = .012$). Also, typical samples' performance was found slightly worse than the atypical samples ($b = -0.152$, $SE = 0.063$, $p = .049$). However, this latter finding should be interpreted with caution because the moderator was barely significant ($p = .049$). These four moderators explained nearly all the true heterogeneity ($\omega^2 = 0.015$, $\tau^2 = 0.001$).

The pattern of results regarding near-transfer effects at follow-up was thus the same as that at immediate post-test: significant overall effects and some true heterogeneity mainly explained by Baseline and Criterion moderators. The only difference was the size of the overall effect. In fact, the post-test overall near-transfer effect was somewhat greater than the follow-up one ($\bar{g} = 0.444$ and $\bar{g} = 0.365$, respectively).

4.2.2.1. Analysis of criterion moderator. Like immediate-post-test effects, the degree of overlap between the training tasks of CWMT and the memory tasks was a significant moderator at follow-up. We thus ran separate analyses for the very-near-transfer measures and lesser-near-transfer measures. First, we ran an RVE model including all the effect sizes related to very-near-transfer measures. The overall effect size was $\bar{g} = 0.487$, $SE = 0.091$, 95% CI [0.292; 0.682], $m = 22$, $k = 60$, $df = 13.27$, $p < .001$, $\omega^2 = 0.017$, $\tau^2 = 0.118$. We ran a meta-regression model including all the moderators. No moderator was significant.

We found one influential case ($g = 1.998$, $ID = 99$) and another effect was excluded because, like in the previous model, it was a

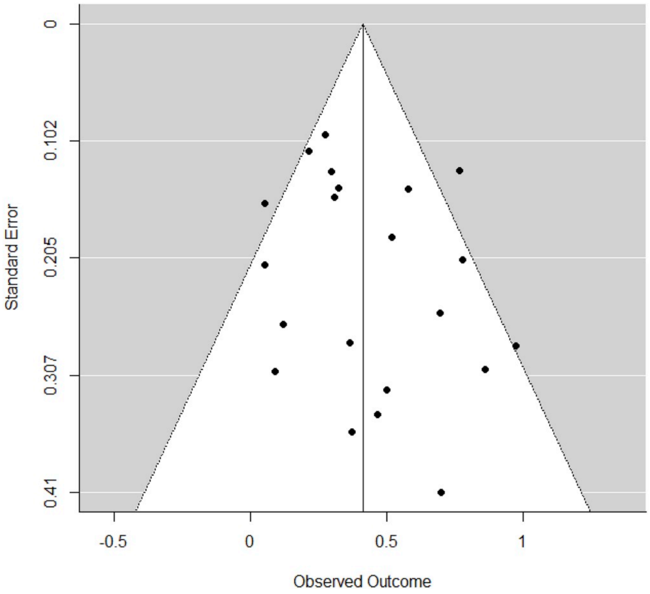

Fig. 6. Funnel plot of observed outcomes (*g*s) and standard errors of very-near-transfer measures at follow-up.

considered an outlier ($g = 3.062$, *ID* = 553). After excluding these effects, the overall effect size was $\bar{g} = 0.445$, $SE = 0.072$, 95% CI [0.290; 0.600], $m = 21$, $k = 58$, $df = 13.48$, $p < .001$, $\omega^2 = 0.047$, $\tau^2 = 0.038$. Baseline was the only significant moderator ($b = -0.451$, $SE = 0.191$, $p = .037$, $\omega^2 = 0.036$, $\tau^2 = 0.017$).

The funnel plot looked slightly asymmetrical (Fig. 6).

The trim-and-fill estimates were $\bar{g} = 0.380$ ($SE = 0.060$, $p < .001$), $\bar{g} = 0.295$ ($SE = 0.068$, $p < .001$), and $\bar{g} = 0.234$ ($SE = 0.068$, $p < .001$) with the *L0*, *R0*, and *Q0* estimators, respectively. The PET estimate was $\bar{g} = 0.239$ ($SE = 0.134$, $p = .090$), and the PEESE estimate was $\bar{g} = 0.326$ ($SE = 0.076$, $p < .001$). As seen above, the PET estimator overcorrected (one-tailed $p < .100$). Thus, the PEESE estimator was more trustworthy.

The results at follow-up were very similar to the ones observed immediately after the post-test. Once again, the only difference was represented by the size of the effects. The uncorrected overall effect at follow-up was slightly smaller than its homologous at immediate post-test ($\bar{g} = 0.487$ and $\bar{g} = 0.566$, respectively). This pattern of results was slightly magnified in the publication-bias-corrected estimates. While the corrected estimates were about $\bar{g} = 0.450$ at immediate post-test, they ranged between $\bar{g} = 0.234$ and $\bar{g} = 0.380$ at follow-up. This difference probably depended on post-test-to-follow-up attrition rate (i.e., only a portion of the studies reported follow-up effects).

Second, we ran an RVE model including all the effect sizes related to lesser-near-transfer measures at follow-up. The overall effect size was $\bar{g} = 0.176$, $SE = 0.049$, 95% CI [0.065; 0.286], $m = 16$, $k = 45$, $df = 9.61$, $p = .005$, $\omega^2 = 0.019$, $\tau^2 = 0.000$. We ran a meta-regression model including all the moderators. Consistent with the very low true heterogeneity observed in the model, no moderator was significant. We found one influential case ($g = -0.235$, *ID* = 551). After excluding this effect, the overall effect size was $\bar{g} = 0.209$, $SE = 0.039$, 95% CI [0.120; 0.299], $m = 15$, $k = 44$, $df = 8.40$, $p < .001$, $\omega^2 = 0.000$, $\tau^2 = 0.000$. Again, since no true heterogeneity was observed, no moderator was significant.

The funnel plot looked slightly asymmetrical (Fig. 7).

The trim-and-fill estimates were $\bar{g} = 0.189$ ($SE = 0.047$, $p < .001$), $\bar{g} = 0.149$ ($SE = 0.043$, $p < .001$), and $\bar{g} = 0.178$ ($SE = 0.046$, $p < .001$) with the *L0*, *R0*, and *Q0* estimators, respectively. The PET estimate was $\bar{g} = 0.089$ ($SE = 0.078$, $p = .270$), and the PEESE estimate was $\bar{g} = 0.154$ ($SE = 0.044$, $p = .004$). Thus, some evidence of publication bias was found. Overall, with the exception of the PET estimator, all the other corrected estimates were, albeit quite small, still significantly different from zero.

Once again, the effect sizes at follow-up were smaller than the ones obtained at immediate post-test. This applied both to the uncorrected estimates ($\bar{g} = 0.171$ and $\bar{g} = 0.248$, respectively) and publication-bias-corrected estimates (approximatively $\bar{g} = 0.150$–0.160 and $\bar{g} = 0.250$–0260, respectively).

5. Discussion

The present paper aimed to analyze the impact of CWMT on people's cognitive function and academic achievement. While the training regimen increased the performance on memory tasks, no appreciable effect was found in far-transfer tasks (no estimate significantly different from zero). The overall effect was estimated to be around zero at follow-up.[6] These outcomes corroborate the

[6] Interestingly, the lack of follow-up far-transfer effects also seems to reject the hypothesis that some time is needed in order for generalized

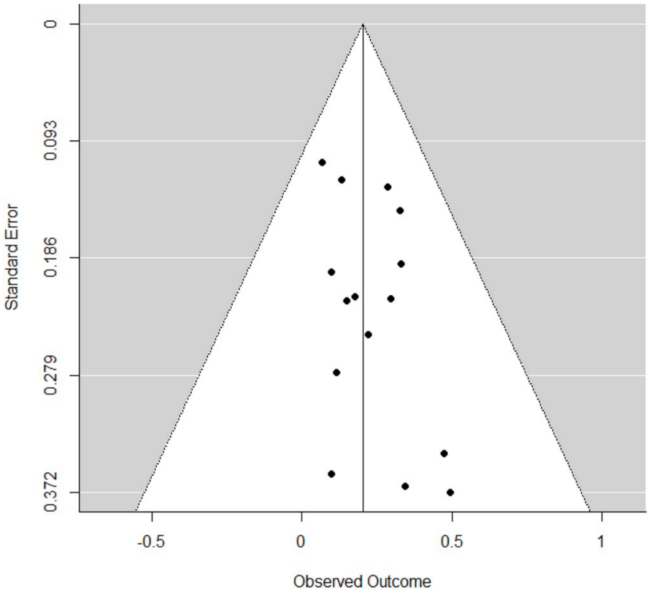

Fig. 7. Funnel plot of observed outcomes (*gs*) and standard errors of lesser-near-transfer measures at follow-up.

results reported in most recent meta-analyses and systematic reviews in the broader field of WM training (e.g., Dougherty, Hamovits, & Tidwell, 2016; Gillam, Holbrook, Mecham, & Weller, 2018; Melby-Lervåg et al., 2016; Sala & Gobet, 2017a, b; Soveri, Antfolk, Karlsson, Salo, & Laine, 2017). Conversely, our findings contradict the more optimistic conclusions of those meta-analyses and systematic reviews specifically examining the impact of CWMT on far-transfer measures (Nutley & Ralph, n.d.; Pearson, 2016; Shinaver et al., 2014; Spencer-Smith & Klingberg, 2015).

The discrepancy between our findings and the ones reported in the latter reviews and meta-analyses stems from several factors. First, and most obvious, our meta-analysis includes many more studies than previous meta-analyses. The inclusion of the most recent studies has increased the precision of the estimated effects and corrected the early optimistic findings. Second, our inclusion criteria are stricter and have led to the exclusion of those studies whose experimental design does not meet a minimum standard of quality (e.g., inclusion of a control group and exclusion of subjective measures of cognitive/academic skill). In fact, poor design quality is often associated with more optimistic results in the field of cognitive training (Simons et al., 2016). Third, unlike the previous meta-analyses, we have employed a set of up-to-date meta-analytic techniques (e.g., RVE) and diagnostics (e.g., influential-case analysis and multiple publication-bias analyses). These methods are necessary to produce unbiased and more reliable estimates (Appelbaum et al., 2018). The quality of the modeling approach may thus explain the difference between ours and the other meta-analyses in the field.

Crucially, the far-transfer meta-analytic models did not exhibit any true between-study or within-study heterogeneity ($\omega^2 = 0.000$ and $\tau^2 = 0.000$ when baseline differences are controlled for). Overall, the results regarding far transfer are consistent with a non-phenomenon: no generalized effects occurred regardless of any potential moderator (e.g., outcome measure, age, or type of population). This consistency is, as far as we are concerned, the most significant novel aspect regarding this particular field of research. In fact, from a statistical point of view at the very least, the results referring to far-transfer effects were not mixed at all. Therefore, the idea that WM-training programs such as CWMT exert stronger benefits to low-WM individuals (e.g., Klingberg et al., 2002; Weicker et al., 2016) is not supported. Overall, these findings thus corroborate the hypothesis according to which the lack of broad generalization of cognitive skills acquired by training is an invariant of human cognition (Sala et al., 2019). To date, the empirical evidence indicates that the possibility of enhancing general cognition by training is scientifically implausible (e.g., Moreau, Macnamara, & Hambrick, 2018; Sala & Gobet, 2019). As pointed out by some scholars (e.g., Engle, 2015), human cognition is the product of a biological system. Thus, it is very unlikely that any short-term cognitive-training program could significantly affect it. CWMT appears to be no exception. Consequently, to date, CWMT cannot be recommended as an educational tool at any age and for any population. Furthermore, since the overall idea of fostering cognitive skills by training seems substantially implausible, these findings cast some doubts about the claimed positive effects of other commercial cognitive-training programs (e.g., Neuroracer, Cognifit, and Lumosity, just to mention some). In this respect, the present meta-analysis is in line with the general skepticism about the alleged benefits of commercial cognitive-training programs expressed by Simon et al. (2016) and reported in large trials (e.g., ACTIVE; Rebok et al., 2014). Future studies will contribute to refute or corroborate this view. Finally, given the consistent lack of broad generalization of skills, it is our conviction that other types of intervention should be preferred in order to improve academic achievement. More promising examples include teaching learning strategies (for a review, see McCabe, Redick, & Engle, 2016) and

(*footnote continued*)
benefits to occur (e.g., Pearson, 2016, p. 17).

increasing the time spent in formal education (Ritchie & Tucker-Drob, 2018).

The near-transfer effects deserve a more nuanced discussion. The training program increased performance on memory tasks immediately after post-test ($\bar{g} = 0.444$) and this improvement remained significant after several months, although it slightly decreased ($\bar{g} = 0.365$ at follow-up). The models showed some amount of within- and between-study true heterogeneity. Most of this heterogeneity was explained by the between-group differences at baseline and similarity between the training tasks in CWMT and memory tasks. As expected, the participants improved the most in those memory tasks whose demands and visual stimuli were very similar to the trained tasks (very-near-transfer; $\bar{g} = 0.566$ and $\bar{g} = 0.487$ immediately after post-test and at follow-up, respectively). These effects appeared to be relatively robust to publication bias (realistically, no more than 0.100–0.150 standardized mean difference of bias). As already highlighted by many researchers in the field (e.g., Shipstead et al., 2012a, b; Simons et al., 2016) and some of the authors of the primary studies included in this meta-analysis (e.g., Brehmer, Westerberg, & Bäckman, 2012), these effects should not be interpreted as evidence of memory enhancement. Rather, such effects denote improvement in the ability to perform the trained tasks.

CWMT also seems to exert a moderate effect on the participants' performance on those memory tasks not included in the training program or related to the trained tasks (lesser-near-transfer). The overall effect sizes were small but significantly different from zero; they remained significantly different from zero for a few months after the end of the training, although some decrease was observed ($\bar{g} = 0.246$, $p = .002$ and $\bar{g} = 0.176$, $p = .005$, at post-test and follow-up, respectively; see also Tables S1–S4 in the Supplemental material available online). These effects were also highly consistent ($\omega^2 = 0.000$ and $\tau^2 \leq 0.001$ after controlling for baseline differences and excluding the few influential cases and outliers). Nevertheless, some evidence of publication bias was found at follow-up (e.g., the PET estimate was $\bar{g} = 0.089$).

5.1. Does CWMT enhance working memory?

The findings regarding far-transfer and very-near-transfer effects are easily interpretable: far-transfer does not exist with CWMT and very-near transfer indicates that the acquired skills can be used in highly similar tasks. In contrast, no straightforward explanation is possible for the improvements on those memory tasks not directly related to CWMT training tasks. A possibility is that the observed lesser-near-transfer effects stem from genuine cognitive enhancement. That is, CWMT may slightly increase WM capacity. The alternative possibility is that CWMT makes the participants more able to perform a certain class of tasks. For example, Shipstead et al. (2012a) have noticed that, even though complex span tasks (e.g., odd one out) are usually categorized as lesser-near-transfer (using our nomenclature), they still share some degree of overlap with the trained tasks (mostly simple-span tasks). Thus, people undergoing CWMT training may simply acquire the ability to perform such tasks slightly more efficiently than controls. That would explain the small observed effect sizes in the near-transfer measures and the concurrent absence of far transfer.

In line with Shipstead et al. (2012a, b), our opinion is that CWMT does not foster WM capacity, any other core cognitive mechanism, or academic skills. Two considerations lead us to this conclusion. First, the effect sizes observed in lesser-near-transfer measures are quite small and tend to diminish a few months after the end of the training. This result can be accounted for by the moderate, yet meaningful, degree of overlap between the trained tasks and memory tasks. Second, and most crucially, WM capacity is a major predictor of academic achievement and is highly correlated with fluid intelligence. Also, as seen earlier, low WM capacity is comorbid with several learning disabilities. Enhanced WM capacity is supposed to make information processing more efficient, which, in turn, should bring a wide set of benefits in academic, professional, and social contexts (see Pearson, 2016). Thus, if CWMT training program were enhancing the participants' WM capacity, improvements in other cognitive and academic tasks should have been observed at either post-test assessment or follow-up assessment. However, this was not the case.

That being said, we think that the topic deserves further investigation. Specifically, the field would substantially benefit from the study of the impact of the training on latent factors rather than observed variables. Cognitive skills are commonly defined as the shared variance between many different cognitive tasks (e.g., Strata II and III of the CHC model; McGrew, 2009). Improvements on a latent factor extracted from a broad set of memory tasks would represent far more compelling evidence of cognitive enhancement than that often provided in the reviewed primary studies, which are based on few observed measures. Such an experimental design would dramatically contribute to settling the debate regarding the true significance of near transfer induced by CWMT and any other cognitive-training program.

5.2. Conclusions

This meta-analysis has examined the impact of CWMT on people's performance on cognitive tests. Small to null effects were observed on far-transfer measures (i.e., fluid intelligence, attention, and mathematical/language skills). The findings were highly consistent (i.e., very low or no true heterogeneity). Thus, the CWMT had no appreciable impact on overall cognitive ability or academic skills.

More robust effects were found on measures of WM capacity very similar to the trained tasks (e.g., digit span and span board tasks). Nevertheless, CWMT exerted only a small effect on measures of WM capacity not directly linked to the trained tasks. Differences at baseline accounted for most of the observed true heterogeneity. Because of the small size of the effects and the lack of generalization across other cognitive and academic skills, the presumed benefits of CWMT on WM capacity remain doubtful. Future studies should test the effect of CWMT on latent factors estimated from many different measures of WM capacity.

Authors' note

NDA and GS share first authorship. All the authors conceptualized this paper. GS performed the statistical analyses. NDA and GS extracted and coded the effect sizes. All the authors contributed to writing the paper.

Acknowledgements

We gratefully thank all the researchers who provided unpublished data. The support of the Japan Society for the Promotion of Science [GS; 17F17313] is gratefully acknowledged. We also thank Walter Boot and Daniel Simons for their feedback on the classification of near-transfer effects.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.edurev.2019.04.003.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin, 131*, 30–60. https://doi.org/10.1037/0033-2909.131.1.30.

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.

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

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*, 612–637. https://doi.org/10.1037/0033-2909.128.4.612.

Brehmer, Y., Westerberg, H., & Bäckman, L. (2012). Working-memory training in younger and older adults: Training gains, transfer, and maintenance. *Frontiers in Human Neuroscience, 6*, 63. https://doi.org/10.3389/fnhum.2012.00063.

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.

Cortese, S., Ferrin, M., Brandeis, D., Buitelaar, J., Daley, D., Dittmann, R. W., ... European ADHD Guidelines Group.. (2015). Cognitive training for Attention-Deficit/Hyperactivity Disorder: Meta-analysis of clinical and neuropsychological outcomes from randomized controlled trials. *Journal of the American Academy of Child & Adolescent Psychiatry, 54*, 164–174. https://doi.org/10.1016/j.jaac.2014.12.010.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*, 13–21. https://doi.org/10.1016/j.intell.2006.02.001.

Detterman, D. K. (2014). Introduction to the intelligence special issue on the development of expertise: Is ability necessary? *Intelligence, 45*, 1–5. https://doi.org/10.1016/j.intell.2014.02.004.

Detterman, D. K. (2016). Education and intelligence: Pity the poor teacher because student characteristics are more significant than teachers or schools. *Spanish Journal of Psychology, 19*, E93. https://doi.org/10.1017/sjp.2016.88.

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 doi:10.1111/desc.12068.

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.

Engle, R. W. (2015). Retrieved from https://www.npr.org/sections/health-shots/2015/08/10/430149726/will-doctors-soon-be-prescribing-video-games-for-mental-health.

Engle, R. W. (2018). Working memory and executive attention: A revisit. *Perspectives on Psychological Science, 13*, 190–193. https://doi.org/10.1177/1745691617720478.

Fisher, Z., Tipton, E., & Zhipeng, H. (2017). *Package "robumeta."*. Retrieved from https://cran.r-project.org/web/packages/robumeta/robumeta.pdf.

Foy, J. G., & Mann, V. (2014). Adaptive cognitive training enhances executive control and visuospatial and verbal working memory in beginning readers. *International Education Research, 2*, 19–43. https://doi.org/10.12735/ier.v2i2p19.

Gillam, S., Holbrook, S., Mecham, J., & Weller, D. (2018). Pull the Andon rope on working memory capacity interventions until we know more. *Language, Speech, and Hearing Services in Schools, 49*, 434–448. https://doi.org/10.1044/2018_LSHSS-17-0121.

Gobet, F. (2016). *Understanding expertise: A multi-disciplinary approach.* London: Palgrave/Macmillan.

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.

Hadwin, J. A., & Richards, H. J. (2016). Working memory training and CBT reduces anxiety symptoms and attentional biases to threat: A preliminary study. *Frontiers in Psychology, 7*, 47. https://doi.org/10.3389/fpsyg.2016.00047.

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.

Honoré, N., & Noël, M. P. (2017). Can working memory training improve preschoolers' numerical abilities? *Journal of Numerical Cognition, 3*, 516–539. https://doi.org/10.5964/jnc.v3i2.54.

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.

Jerrim, J., Macmillan, L., Micklewright, J., Sawtell, M., & Wiggins, M. (2016). *Chess in Schools. Evaluation report and rxecutive rummary*Education Endowment Foundation. Retrieved from: https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/EEF_Project_Report_Chess_in_Schools.pdf.

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.

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., Forssberg, H., & Westerberg, H. (2002). Training of working memory in children with ADHD. *Journal of Clinical and Experimental Neuropsychology, 24*, 781–791. https://doi.org/10.1076/jcen.24.6.781.8395.

Libertus, M. E., Liu, A., Pikul, O., Jacques, T., Cardoso-Leite, P., Halberda, J., et al. (2017). The impact of action video game training on mathematical abilities in adults. *AERA Open, 3*, 1–13. https://doi.org/10.1177/2332858417740857.

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.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*, 1–10. https://doi.org/10.1016/j.intell.2008.08.004.

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., 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.

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 sciences.* Advanced online publicationhttps://doi.org/10.1177/0963721418797300.

Nutley, S., & Ralph, K. Cogmed: Published research meta-analysis. Retrieved from https://www.pearsonclinical.co.uk/Cogmed/Downloads/cogmed-research-meta-analysis.pdf.

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.

Pearson (2016). Cogmed working memory training: Claims & evidence – extended version V.4.1. Retrieved from https://www.cogmed.com/wp-content/uploads/CogmedClaimsEvidence.pdf.

Peng, P., Barnes, M., Wang, C., Wang, W., Li, S., Swanson, H. L., et al. (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.

Rapport, M. D., Orban, S. A., Kofler, M. J., & Friedman, L. M. (2013). Do programs designed to train working memory, other executive functions, and attention benefit children with ADHD? A meta-analytic review of cognitive, academic, and behavioral outcomes. *Clinical Psychology Review, 33*, 1237–1252. https://doi.org/10.1016/j.cpr.2013.08.005.

Rebok, G. W., Ball, K., Guey, L. T., Jones, R. N., Kim, H.-Y., King, J. W., et al. (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.

Rindermann, H., & Neubauer, A. C. (2004). Processing speed, intelligence, creativity, and school performance: Testing of causal hypotheses using structural equation models. *Intelligence, 32*, 573–589. https://doi.org/10.1016/j.intell.2004.06.005.

Ritchie, S. J., & Tucker-Drob, E. M. (2018). How much does education improve intelligence? A meta-analysis. *Psychological Science, 29*, 1358–1369. https://doi.org/10.1177/095679761877425.

Roberts, G., Quach, J., Spencer-Smith, M., Anderson, P., Gathercole, S., Gold, L., et al. (2016). Academic outcomes 2 years after working memory training for children with low working memory: A randomized clinical trial. *JAMA Pediatrics, 170*, e154568. https://doi.org/10.1001/jamapediatrics.2015.4568.

Roche, J. D., & Johnson, B. D. (2014). Cogmed working memory training product review. *Journal of Attention Disorders, 18*, 379–384. https://doi.org/10.1177/1087054714524275.

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. (2017a). 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. (2017b). Does far transfer exist? Negative evidence from chess, music, and working memory training. *Current Directions in Psychological Science, 26*, 515–520. https://doi.org/10.1177/0963721417712760.

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.

Schellenberg, E. G. (2004). Music lessons enhance IQ. *Psychological Science, 15*, 511–514. https://doi.org/10.1111/j.0956-7976.2004.00711.x.

Schmidt, F. L. (2017). Beyond questionable research methods: The role of omitted relevant research in the credibility of research. *Archives of Scientific Psychology, 5*, 32–41. https://doi.org/10.1037/arc0000033.

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Newbury Park, CA: Sage.

SharpBrains (2015, January). *The digital brain health market 2012-2020: Web-based, mobile and biometrics-based technology to assess, monitor, and enhance cognition and brain functioning [Addendum].* San Francisco, CA: SharpBrains.

Shinaver, C. S., III, Entwistle, P. C., & Söderqvist, S. (2014). Cogmed WM training: Reviewing the reviews. *Applied Neuropsychology: Child, 3*, 163–172. https://doi.org/10.1080/21622965.2013.875314.

Shipstead, Z., Hicks, K. L., & Engle, R. W. (2012a). Cogmed working memory training: Does the evidence support the claims? *Journal of Applied Research in Memory and Cognition, 1*, 185–193. https://doi.org/10.1016/j.jarmac.2012.06.003.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012b). 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., et al. (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.

Sonuga-Barke, E. J., Brandeis, D., Cortese, S., Daley, D., Ferrin, M., Holtmann, M., ... European ADHD Guidelines Group.. (2013). Nonpharmacological interventions for ADHD: Systematic review and meta-analyses of randomized controlled trials of dietary and psychological treatments. *American Journal of Psychiatry, 170*, 275–289. https://doi.org/10.1176/appi.ajp.2012.12070991.

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.

Spencer-Smith, M., & Klingberg, T. (2015). Benefits of a working memory training program for inattention in daily life: A systematic review and meta-analysis. *PLoS One, 10*, e0119522. https://doi.org/10.1371/journal.pone.0119522.

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.

Strobach, T., & Karbach, J. (2016). *Cognitive training: An overview of features and applications.* New York: Springer.

Süß, H. M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability - and a little bit more. *Intelligence, 30*, 261–288. https://doi.org/10.1016/S0160-2896(01)00100-3.

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. (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: Springer.

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods, 5*, 13–30. https://doi.org/10.1002/jrsm.1091.

Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling complex meta-analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental and Life-Course Criminology, 2*, 85–112. https://doi.org/10.1007/s40865-016-0026-5.

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.

Villemonteix, T. (2018). L'entraînement de la mémoire de travail est-il bénéfique pour les enfants présentant un trouble déficit de l'attention/hyperactivité? *Neuropsychiatrie de l'Enfance et de l'Adolescence, 66*, 3–12. https://doi.org/10.1016/j.neurenf.2017.07.003.

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.

Westerberg, H., Hirvikoski, T., Forssberg, H., & Klingberg, T. (2004). Visuo-spatial working memory span: A sensitive measure of cognitive deficits in children with ADHD. *Child Neuropsychology, 10*, 155–161. https://doi.org/10.1080/09297040490911014.