Journal of Cognitive Enhancement (2020) 4:192–210
https://doi.org/10.1007/s41465-020-00164-6

REVIEW

Quantifying the Difference Between Active and Passive Control Groups in Cognitive Interventions Using Two Meta-analytical Approaches

Jacky Au[1] · Benjamin C. Gibson[2] · Kimberly Bunarjo[3,4] · Martin Buschkuehl[3] · Susanne M. Jaeggi[1,4]

Received: 22 August 2019 / Accepted: 13 January 2020 / Published online: 29 January 2020

Abstract

Despite promising reports of broad cognitive benefit in studies of cognitive training, it has been argued that the reliance of many studies on no-intervention control groups (passive controls) make these reports difficult to interpret because placebo effects cannot be ruled out. Although researchers have recently been trying to incorporate more active controls, in which participants engage in an alternate intervention, previous work has been contentious as to whether this actually yields meaningfully different results. To better understand the influence of passive and active control groups on cognitive interventions, we conducted two meta-analyses to estimate their relative effect sizes. While the first one broadly surveyed the literature by compiling data from 34 meta-analyses, the second one synthesized data from 42 empirical studies that simultaneously employed both types of controls. Both analyses showed no meaningful performance difference between passive and active controls, suggesting that current active control placebo paradigms might not be appropriately designed to reliably capture these non-specific effects or that these effects are minimal in this literature.

Keywords Placebo effects · Hawthorne effects · Experimental confounds · Cognitive training · Meta-analysis

Being able to make causal claims is a primary goal of experimental scientists. A strong demonstration of causality usually entails both a clear temporal relationship between two variables (i.e., cause precedes effect), as well as the manipulation and isolation of a single causal factor (i.e., two comparison groups that are completely matched on all variables save the one of interest). The former criterion is generally easy to satisfy in experimental studies, but the latter can be more elusive. In the medical field, clinical drug trials can circumvent or mitigate this issue by designing placebo pills meant to provide an identical patient experience as the experimental drug, but which is missing the key active ingredient hypothesized to produce therapeutic benefits. These placebos are generally effective at controlling psychological components of treatments by providing a similar clinical context and inducing similar treatment expectations, but may sometimes fall short, such as when obvious side effects occur that clue patients in to their assigned treatment arm. Things can get even more problematic in other fields of research, however, when the precise therapeutic ingredient has not been well-elucidated and thus cannot be effectively isolated. Or even if it has been, the intervention properties may be more difficult to disentangle from non-specific effects, especially if the therapeutic ingredients themselves have a strong psychological component (e.g., Kirsch 2005). In the following, we focus on the cognitive training field, where both these limitations exist.

Cognitive training—or often colloquially referred to as “brain training”—encompasses a broad field of research, whereby the primary goal is to enhance certain cognitive skills through behavioral interventions designed to target those or related skills. A pertinent and popular example is working

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s41465-020-00164-6) contains supplementary material, which is available to authorized users.

✉ Jacky Au
jwau@uci.edu

[1] Department of Cognitive Sciences, University of California, Irvine, Irvine, CA 92617, USA

[2] Department of Psychology, University of New Mexico, Albuquerque, NM 87131, USA

[3] MIND Research Institute, Irvine, CA 92617, USA

[4] School of Education, University of California, Irvine, Irvine, CA 92697, USA

memory training, where participants train on memory tasks and/or games that require the simultaneous maintenance and manipulation of multiple pieces of information. One of the primary aims of such working memory-based interventions is not only to improve the specific skills acquired through practicing the training task itself but also more importantly to generalize or "transfer" those skills to other tasks or domains that go beyond the trained task (Pahor et al. 2018). Researchers commonly distinguish between "near" transfer, where the training task and outcome measure presumably share many overlapping features and processes, and "far" transfer, where the outcome measure is thought to be more different from the training task, although it has been difficult to quantify the boundary conditions of such transfer effects (Barnett and Ceci 2002). Nonetheless, various meta-analyses within the working memory training field have demonstrated that near transfer effects are more consistently observed as compared to far transfer effects, and the effect sizes seem to be larger in near transfer measures (e.g., Soveri et al. 2017; Weicker et al. 2016). Neuroimaging work has been consistent with that distinction in that the frontostriatal system seems to mediate near transfer effects, whereas the dorso- and ventrolateral prefrontal cortices seem to engage domain-general networks that facilitate learning more broadly (e.g., Salmi et al. 2018, for a recent meta-analysis).

Even though the causal rationale of transfer effects in working memory training (i.e., training working memory leads to improvements in tasks that rely on working memory) seems straightforward at first, the task impurity problem (e.g., Schweizer 2007) and the fact that it is quite common that different working memory tasks often share less than 25% of variance (e.g., Ackerman et al. 2005) complicates this rationale considerably. Therefore, it is not surprising that, unlike in pharmaceutical studies, it is often difficult in behavioral interventions to isolate one single component theorized to confer a cognitive benefit. It is a non-trivial and perhaps even impossible endeavor to strip a working memory training task of its working memory component and still maintain the perceived integrity of the intervention the same way a pill can be rendered inert by replacing its active ingredients with sugar yet still remain believable to participants. Early cognitive training studies often took a broad strokes approach and simply compared intervention groups to no-intervention control groups, an approach which is commonplace in educational contexts where the controls often consist of "business as usual" groups, or in the clinical field, where "wait-list controls" are employed in order to provide everyone the chance to eventually receive the potentially effective intervention. These no-intervention controls, hereafter referred to as passive control groups, effectively control for practice effects on the outcome measures by completing pre- and posttest assessments, but having minimal or no contact with experimenters or any part of the experimental protocol in the interim. Early studies in the cognitive training field using these types of designs were very influential (Chein and Morrison 2010; Dahlin et al. 2008; Jaeggi et al. 2008; Schmiedek et al. 2010), but arguably might have overestimated the actual impacts of cognitive training. Specifically, the use of passive controls exposed these studies to a variety of threats to internal validity related to non-specific characteristics of the training protocol such as experimenter contact, participant and/or experimenter expectancies, and demand characteristics where participants unconsciously conform to what they believe to be the purpose of the experiment (Boot et al. 2013; Nichols and Maner 2008). Therefore, any performance advantage of the experimental group over these passive controls could not be exclusively attributed to the intervention itself since these non-specific characteristics also differed between groups. The advantages and disadvantages of various control groups, as well as the role of beliefs and expectations, are further discussed below, but are also the subject of ongoing debates (Au et al. 2016; Boot et al. 2013; Melby-Lervåg and Hulme 2013; Melby-Lervåg et al. 2016; Shipstead et al. 2012; Tsai et al. 2018).

The presumed solution to these threats adopted by many in the research community has been to use active controls, in which the control group participates in an alternative intervention not designed to target the core cognitive skills of interest. Critics have contended that the use of such active control groups renders the most promising results of cognitive training studies null and have demonstrated via meta-analysis that positive far transfer effects within the sub-field of working memory training are only driven by studies with passive controls (Melby-Lervåg et al. 2016). It has been concluded by many, therefore, that any effects observed as a result of working memory interventions merely reflect placebo and other non-specific artifacts. However, this conclusion seems premature, as the meta-analytic work on which it is based is merely correlational in nature. It cannot be precluded that factors beyond the nature of the control groups themselves contribute to the lack of effects associated with studies that use active controls. Supporting this notion is the finding that the reason working memory training studies with active controls yield smaller meta-analytic effect sizes than studies with passive controls is due to differences in performance of the *experimental* groups between the two types of studies, rather than differences between the control groups. In fact, when looking at the pre-post changes *within* the control groups, both passive and active controls perform pretty similarly (Au et al. Au et al. 2015, 2016 see Fig. 1b; Soveri et al. 2017). Though it is not known what might cause this discrepant performance among the experimental groups between these two types of studies, there is currently no direct evidence within the subset of studies analyzed to date supporting the claim that positive training effects only arise as a result of using passive controls (Au et al. 2016).

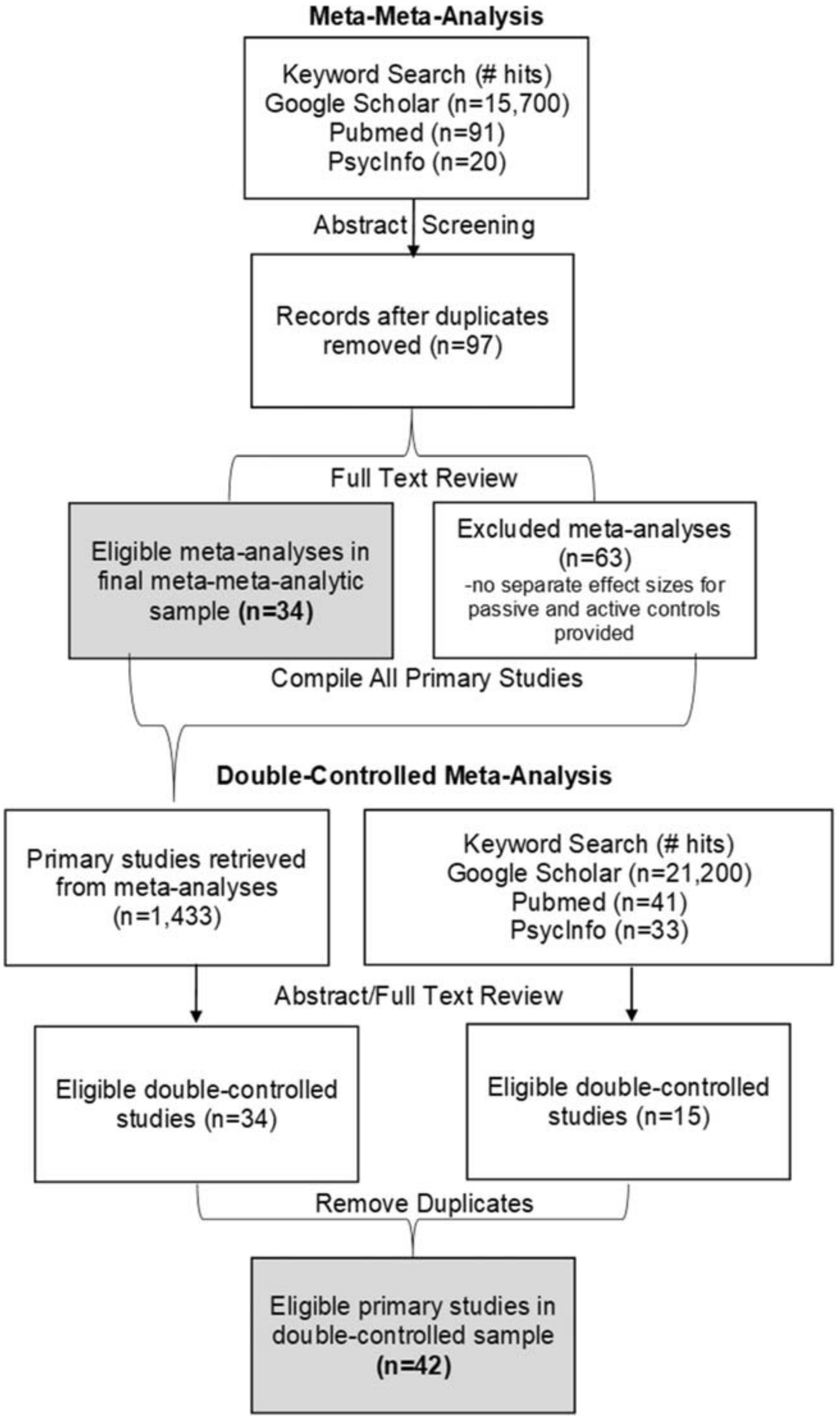

Meta-Meta-Analysis
Keyword Search (# hits)
Google Scholar (n=15,700)
Pubmed (n=91)
PsycInfo (n=20)
Abstract Screening
Records after duplicates removed (n=97)
Full Text Review
Eligible meta-analyses in final meta-meta-analytic sample (n=34)
Excluded meta-analyses (n=63)
-no separate effect sizes for passive and active controls provided
Compile All Primary Studies
Double-Controlled Meta-Analysis
Primary studies retrieved from meta-analyses (n=1,433)
Keyword Search (# hits)
Google Scholar (n=21,200)
Pubmed (n=41)
PsycInfo (n=33)
Abstract/Full Text Review
Eligible double-controlled studies (n=34)
Eligible double-controlled studies (n=15)
Remove Duplicates
Eligible primary studies in double-controlled sample (n=42)

Fig. 1 Flow chart of study extraction process

Nevertheless, the popular belief among researchers remains that interpretation of the cognitive training field should rely solely on studies that use active controls, while discounting studies that use passive controls (Melby-Lervåg et al. 2016). The current work seeks to contribute data to this debate by meta-analyzing studies from an extensive range of cognitive interventions going beyond just the field of working memory training in order to comprehensively quantify the performance difference between passive and active controls. Using two complementary quantitative approaches, we first performed a meta-meta-analysis, that is, a meta-analysis of existing cognitive training meta-analyses comparing passive and active control groups across studies. This technique has been used before (e.g., Cleophas and Zwinderman 2017) and is an effective way of overviewing a very broad swath of literature. We followed up the meta-meta-analysis with a more direct, but less comprehensive (due to smaller sample size), meta-analysis of primary studies that used both a passive and an active control within the same study. By directly controlling for all other within-study variables that may influence effect sizes, this second meta-analysis goes beyond the correlational findings of the meta-meta-analysis and approximates a causal framework. In both analyses, we hypothesized no meaningful differences between performance of passive and active controls on measures of cognitive function, as we and others have previously observed among working memory training studies (Au et al. 2015, 2016; Soveri et al. 2017).

Methods

Design

We conducted two meta-analyses in order to survey the cognitive training literature and summarize the effect size differences between passive control groups and active control groups. First, we conducted a meta-meta-analysis that broadly surveyed the literature by synthesizing results from 34 cognitive training meta-analyses, which together summarized the effects of more than 1000 primary studies. We followed this up with a traditional meta-analysis of 42 primary studies which all employed both passive and active control groups within the same study (going forward referred to as double-controlled meta-analysis). Both analyses, where possible, adhered to PRISMA guidelines (Moher 2009). Details of these two approaches are further described in their respective sections.

Study Selection and Inclusion Criteria

Figure 1 represents a flow chart of the study extraction process. For both the meta-meta-analysis as well as the double-controlled meta-analysis, we used liberal inclusion criteria in that we attempted to include a comprehensive set of all cognitive training studies reported in English, excluding mixed-intervention studies such as combining a cognitive intervention with electrical brain stimulation or with a physical exercise regimen. For the meta-meta-analysis, we searched the PubMed, PsycInfo, and Google Scholar databases for articles using the following keywords and boolean operators: ("meta-analysis" OR "systematic review" OR "quantitative review") AND ("cognitive training" OR "working memory training" OR "video game training" OR "cognitive remediation"). PubMed returned 91 hits; PsycInfo returned 20 hits; and Google Scholar returned 15,700 hits. For the Google Scholar database, we restricted our search to the first 1000 hits in order to select the most pertinent articles. Abstracts were screened for inclusion and the full text was scanned, if necessary, to make sure the meta-analysis provided enough data to obtain separate effect size estimates for studies with passive and active controls. If not, authors were emailed to provide the relevant information. All meta-analyses that fit these criteria and that were published through the end of 2016 were included in the meta-meta-analysis. In total, 97 meta-analyses were extracted from the literature. The majority of these meta-analyses did not separately report their effect size estimates as a function of control group type (passive vs. active) or did not provide enough information for us to calculate separate effect size and variance estimates on our own and were thus excluded. In the end, 34 meta-analyses provided enough information to be included in the final analysis.

In order to find empirical studies for the double-controlled meta-analysis, we searched through the primary papers listed in all 97 meta-analyses returned from our original search and included any study that contained both a passive and an active control group. Although there were several instances in which authors defined an intervention group to be an active control despite other researchers (including ourselves) using an identical or similar intervention as an experimental training group (Boot et al. 2008; Opitz et al. 2014; Stephenson and Halpern 2013; Thompson et al. 2013; Vartanian et al. 2016), we decided to rely on the authors' characterization of an active control group in all instances in order to reduce the number of subjective decisions made on our part. In total, 1433 articles were searched, and 34 met inclusion criteria. Additionally, in order to supplement this search method, we also used a keyword search with the following keywords and boolean operators: ("placebo training" OR "active placebo control" OR "active control" OR "treated control" OR "training control") AND ("nonactive control" OR "no-contact control" OR "wait-list control" OR "untreated control" OR "nontreated control" OR "passive control") AND ("cognitive training" OR "working memory training" OR "cognitive rehabilitation" OR "cognitive remediation" OR "videogame training" OR "intervention"). PubMed, PsycInfo, and Google Scholar returned 41, 33, and 21,200 hits, respectively, from which we extracted 15 additional eligible studies. See Fig. 1 for a flowchart of the study extraction process, and the Supplemental Online Materials for a complete bibliography of all included studies.

Coding

After study selection was completed, every article was independently coded by at least two members of the author team and answers were automatically compared using an Excel spreadsheet algorithm. Percent agreement was extremely high between authors (>99%) because we made it a point to only code clear, objective variables that require minimal decision-making in order to promote transparency and enhance replicability of our analyses. Disagreements, few as they were, were resolved by group discussion. Effect sizes or test scores that were only available as figures and not as tables were extracted using Webplot Digitizer (Rohatgi 2017).

Statistical Analyses

Effect Size and Bayes Factor Calculations

All effect size calculations were made with the Comprehensive Meta-Analysis (CMA) software package (Borenstein et al. 2005). Effect sizes were weighted by their inverse variance, or precision, and subsequently pooled together using a random effects model (Riley et al. 2011). For both the meta-meta-analysis as well as the double-controlled meta-analysis, three different summary effects were calculated. First, we summarized the effect size of experimental interventions versus passive controls, then the effect size of experimental interventions versus active controls, and finally, we directly compared the effect size of active controls with that of passive controls. Further description of these methods and calculations are detailed below separately for the meta-meta-analysis and the double-controlled meta-analysis.

Since we hypothesized no differences between passive and active control groups, we also implemented Bayesian analyses to quantify the strength of evidence in favor of the null. Bayes factors were calculated using the meta.ttestBF function in the Bayes Factor package in R (Morey et al. 2014; R Core Team 2013) by converting each effect size into its corresponding *Z*-score as reported by CMA (Borenstein 2009). A weighted sum of *Z*-scores following this method provides largely similar results to the inverse variance-weighted average approach described above (Lee et al. 2016). In accordance with Rouder et al. (2009), we set the rscale parameter to 1 to yield a standard Cauchy prior centered on zero. This approach allowed us to take an uninformed "objective" approach that does not rely on a subjective analysis of the prior literature, but allows the prior distribution of true effect sizes to range from negative infinity to positive infinity, with 50% of the probability mass ranging from $d = -1$ to $d = +1$. However, since Bayesian statistics can be strongly influenced by prior selection, we ran a sensitivity analysis with a range of other possible prior specifications ($r = 0.01$, $r = 0.1$, and $r = 0.3$) to represent very small, small, and moderate effect size distributions. The resulting Bayes factors are reported as BF_{10} or BF_{01} to represent evidence supporting either the alternative, or null, hypotheses, respectively. Typically, the evidential value of Bayes factors below 3 are considered weak, between 3 and 10 are considered substantial, between 10 and 30 are considered strong, between 30 and 100 are considered very strong, and over 100 are considered decisive (Jarosz and Wiley 2014).

Meta-meta-analysis

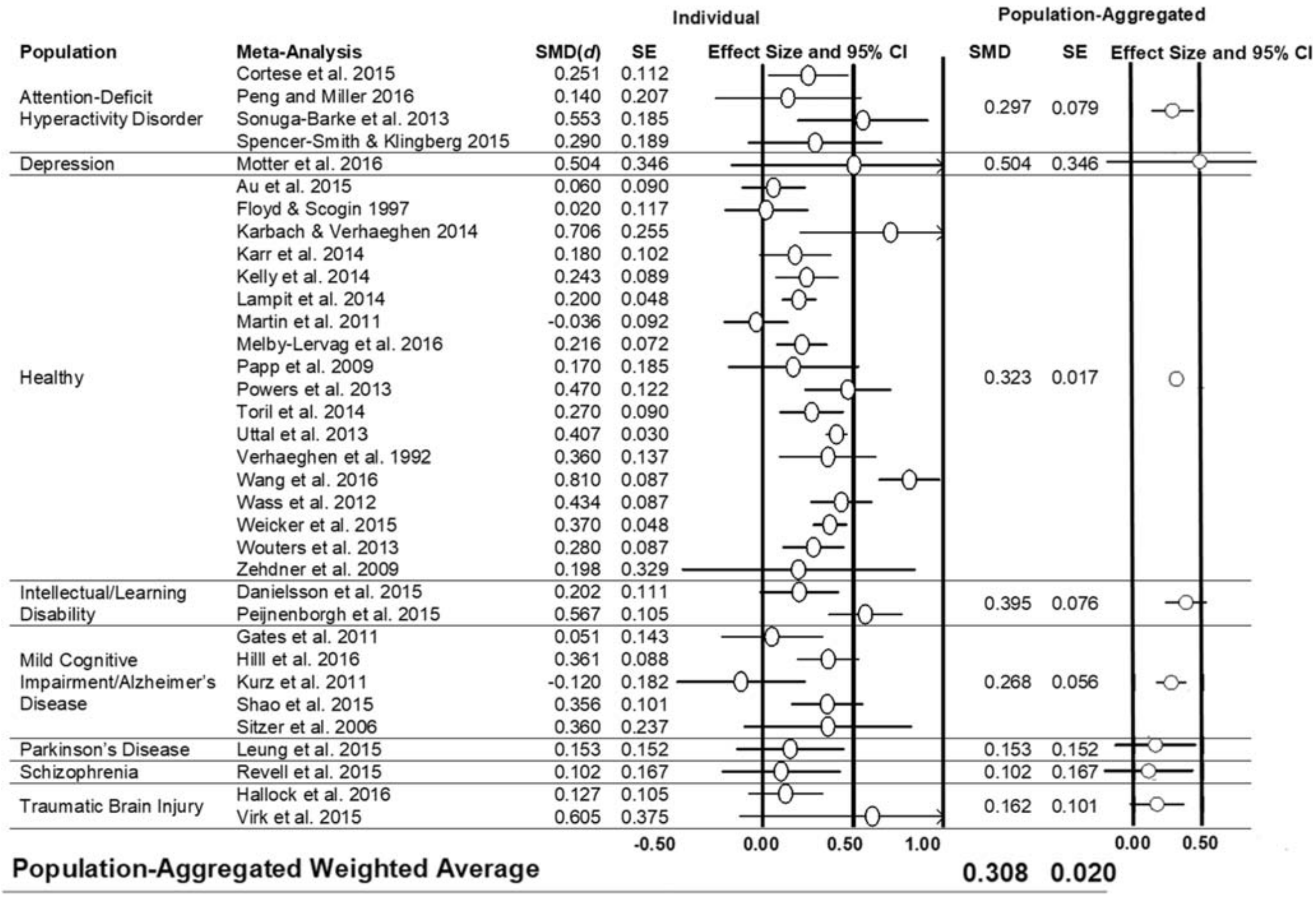
Three different summary effects were calculated, one each for the experimental vs. passive control, experimental vs. active control, and experimental/passive minus experimental/active comparisons. Towards this end, effect sizes were calculated first at the meta-analytic level, then at the population level, and finally at the summary meta-meta-analytic level (see Figs. 2, 3, and 4).

At the meta-analytic level, effect sizes were extracted directly from the individual meta-analyses when provided. All effect sizes represented the standardized mean difference (SMD) in performance between experimental groups and their respective control groups, captured as Cohen's *d*. These effect sizes largely represented performance on laboratory measures of cognitive functioning or self/proxy-reports of behavioral/cognitive improvement, but each meta-analysis had its own criteria for effect size calculations; thus, we simply extracted whatever effect size was reported. When a moderator analysis based on control group type was not directly provided by a particular meta-analysis, we used CMA to calculate the meta-analytic effect sizes, based on the effect sizes reported for individual studies.

However, since some meta-analyses contain overlapping primary studies, we could not simply average the meta-analytic effects to arrive at a summary meta-meta-analytic effect size without violating the assumption of independence. In order to mitigate this issue, we separated the 34 meta-analyses into 8 non-overlapping population categories. We then aggregated effect sizes within each population category in order to come up with 8 distinct and statistically independent effect sizes in which the same study is never represented more than once: attention-deficit/hyperactivity disorder (ADHD; $k = 6$), clinical depression ($k = 1$), healthy individuals ($k = 19$), intellectual or learning disability ($k = 2$), mild cognitive impairment or dementia ($k = 5$), Parkinson's disease ($k = 1$), schizophrenia ($k = 2$), and traumatic brain injury ($k = 2$). This approach allowed us to extract a population-aggregated unit of analysis for the final meta-meta-analysis that does not violate the assumption of independence.[1]

In this way, summary effects at the meta-meta-analytic level were calculated separately for studies with either control type. However, in order to directly compare the relative performance of passive and active controls, a further analysis was conducted following the same procedure as above, but which subtracted out the experimental effects within each meta-analysis. This was done by subtracting the experimental/active control effect size from the experimental/passive control effect size *within* a meta-analysis, while pooling their standard errors together according to the formula: $\mathrm{SE} = \frac{1}{\sqrt{\frac{P_p+P_a}{2}}}$, where P_p = precision (inverse variance) of the experimental/passive control effect size and P_a = precision of the experimental/active control effect size (Borenstein et al. 2005). This left us with a summary meta-analytic effect size capturing the difference in performance between studies with passive controls and studies with active controls. As done above, these summary meta-

[1] Note that none of the clinical studies contained any healthy controls, so all population groups are indeed mutually exclusive.



Population	Meta-Analysis	SMD(*d*)	SE	Population-Aggregated SMD	Population-Aggregated SE
Attention-Deficit Hyperactivity Disorder	Cortese et al. 2015	0.251	0.112	0.297	0.079
	Peng and Miller 2016	0.140	0.207		
	Sonuga-Barke et al. 2013	0.553	0.185		
	Spencer-Smith & Klingberg 2015	0.290	0.189		
Depression	Motter et al. 2016	0.504	0.346	0.504	0.346
Healthy	Au et al. 2015	0.060	0.090	0.323	0.017
	Floyd & Scogin 1997	0.020	0.117		
	Karbach & Verhaeghen 2014	0.706	0.255		
	Karr et al. 2014	0.180	0.102		
	Kelly et al. 2014	0.243	0.089		
	Lampit et al. 2014	0.200	0.048		
	Martin et al. 2011	-0.036	0.092		
	Melby-Lervag et al. 2016	0.216	0.072		
	Papp et al. 2009	0.170	0.185		
	Powers et al. 2013	0.470	0.122		
	Toril et al. 2014	0.270	0.090		
	Uttal et al. 2013	0.407	0.030		
	Verhaeghen et al. 1992	0.360	0.137		
	Wang et al. 2016	0.810	0.087		
	Wass et al. 2012	0.434	0.087		
	Weicker et al. 2015	0.370	0.048		
	Wouters et al. 2013	0.280	0.087		
	Zehdner et al. 2009	0.198	0.329		
Intellectual/Learning Disability	Danielsson et al. 2015	0.202	0.111	0.395	0.076
	Peijnenborgh et al. 2015	0.567	0.105		
Mild Cognitive Impairment/Alzheimer's Disease	Gates et al. 2011	0.051	0.143	0.268	0.056
	Hill et al. 2016	0.361	0.088		
	Kurz et al. 2011	-0.120	0.182		
	Shao et al. 2015	0.356	0.101		
	Sitzer et al. 2006	0.360	0.237		
Parkinson's Disease	Leung et al. 2015	0.153	0.152	0.153	0.152
Schizophrenia	Revell et al. 2015	0.102	0.167	0.102	0.167
Traumatic Brain Injury	Hallock et al. 2016	0.127	0.105	0.162	0.101
	Virk et al. 2015	0.605	0.375		
Population-Aggregated Weighted Average				**0.308**	**0.020**

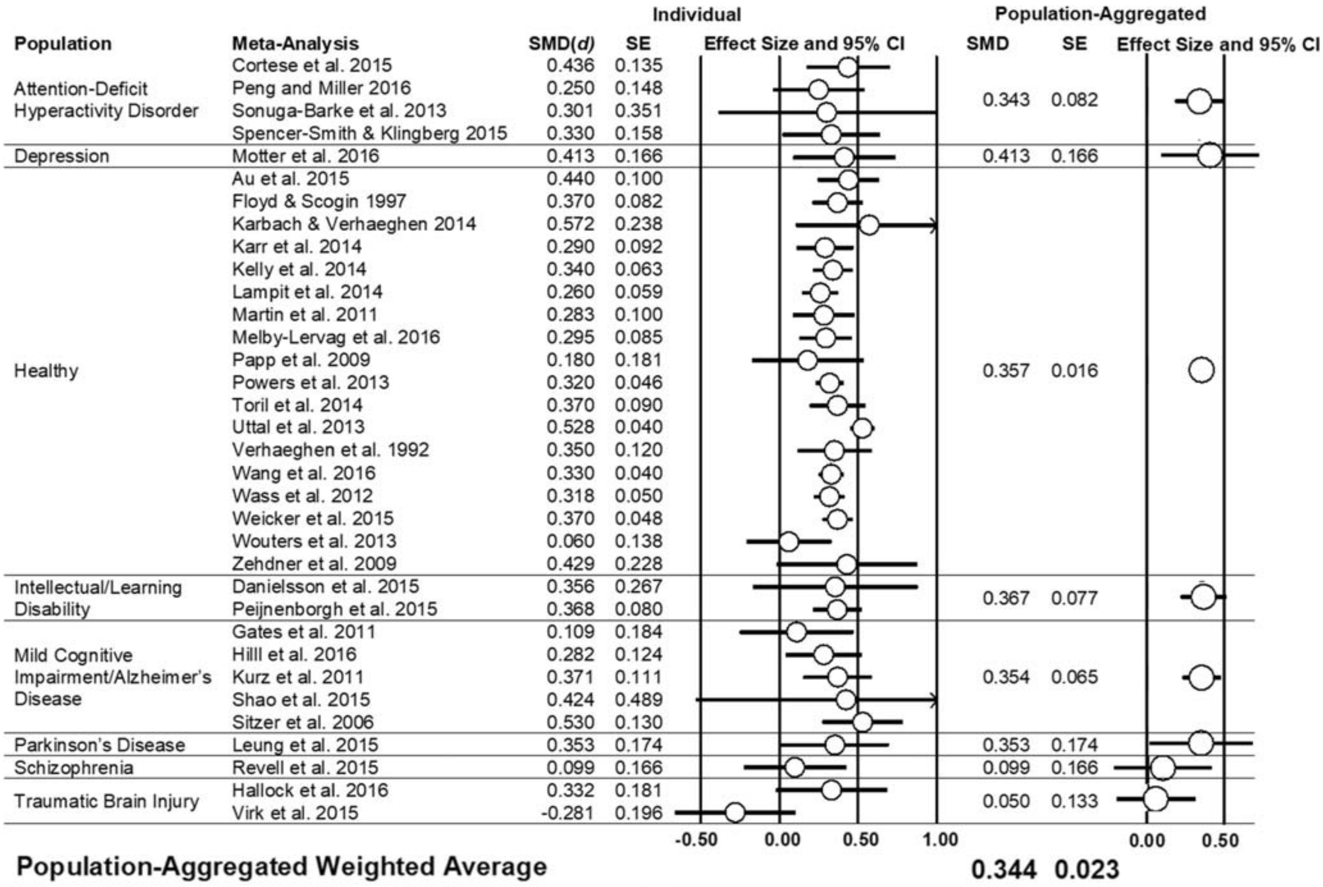
Fig. 2 Meta-meta-analysis of experimental vs. active control comparisons. Across meta-analyses, cognitive training studies using active control groups yield an overall effect size of $d = 0.308$. Positive effect sizes favor experimental groups

analytic effect sizes were then aggregated into population-level effect sizes and then finally into a meta-meta-analytic effect size. It was necessary to conduct this analysis in this paired, within-meta-analysis fashion because effect sizes within the same meta-analysis, whether they are derived from passive or active controls, can be correlated with each other due to idiosyncratic decisions specific to each meta-analysis, such as the choice of outcome measures, inclusion/exclusion criteria, or the method of effect size calculation.

Double-Controlled Meta-analysis

We aggregated data from 42 articles containing 44 independent comparisons between experimental intervention groups, passive control groups, and active control groups. Once again, three summary meta-analytic effect sizes were calculated—one for the experimental/passive control comparison, one for the experimental/active control comparison, and one that directly compares the within-study performance difference of active and passive controls. Only objective cognitive outcome measures that were not specifically trained were included in the calculation of effect sizes. Therefore, subjective or non-cognitive outcomes such as questionnaires, physiological indices, neuroimaging metrics, etc., were excluded. Also, any outcome that was specifically trained by either the experimental or active control group was excluded, such as instances where a group trained on an n-back task and was then tested on the same or a similar n-back task. Thus, our meta-analysis is only focused on transfer to untrained tasks, and not specific training effects.

All studies used a pretest-intervention-posttest design and effect sizes were calculated as the SMD in performance between the groups of interest, after adjusting for small sample sizes using Hedge's *g* (Rosenthal 1991). This was calculated as the mean difference in gain scores on all objective outcome measures within a study, standardized by the pooled standard deviation at pretest (Morris 2008), as used in prior analyses (see Au et al. 2016; Melby-Lervåg and Hulme 2016): $g = \frac{(\mu_{1\mathrm{post}}-\mu_{1\mathrm{pre}})-(\mu_{2\mathrm{post}}-\mu_{2\mathrm{pre}})}{SD_{\mathrm{pre-pooled}}} \times 1-(\frac{3}{4*df-1})$. Effect sizes from all outcomes within a study were averaged together into one net effect, weighted by their inverse variance. Positive values reflect superior pre-post gains among the experimental groups, or the active control group in the case of the active/passive



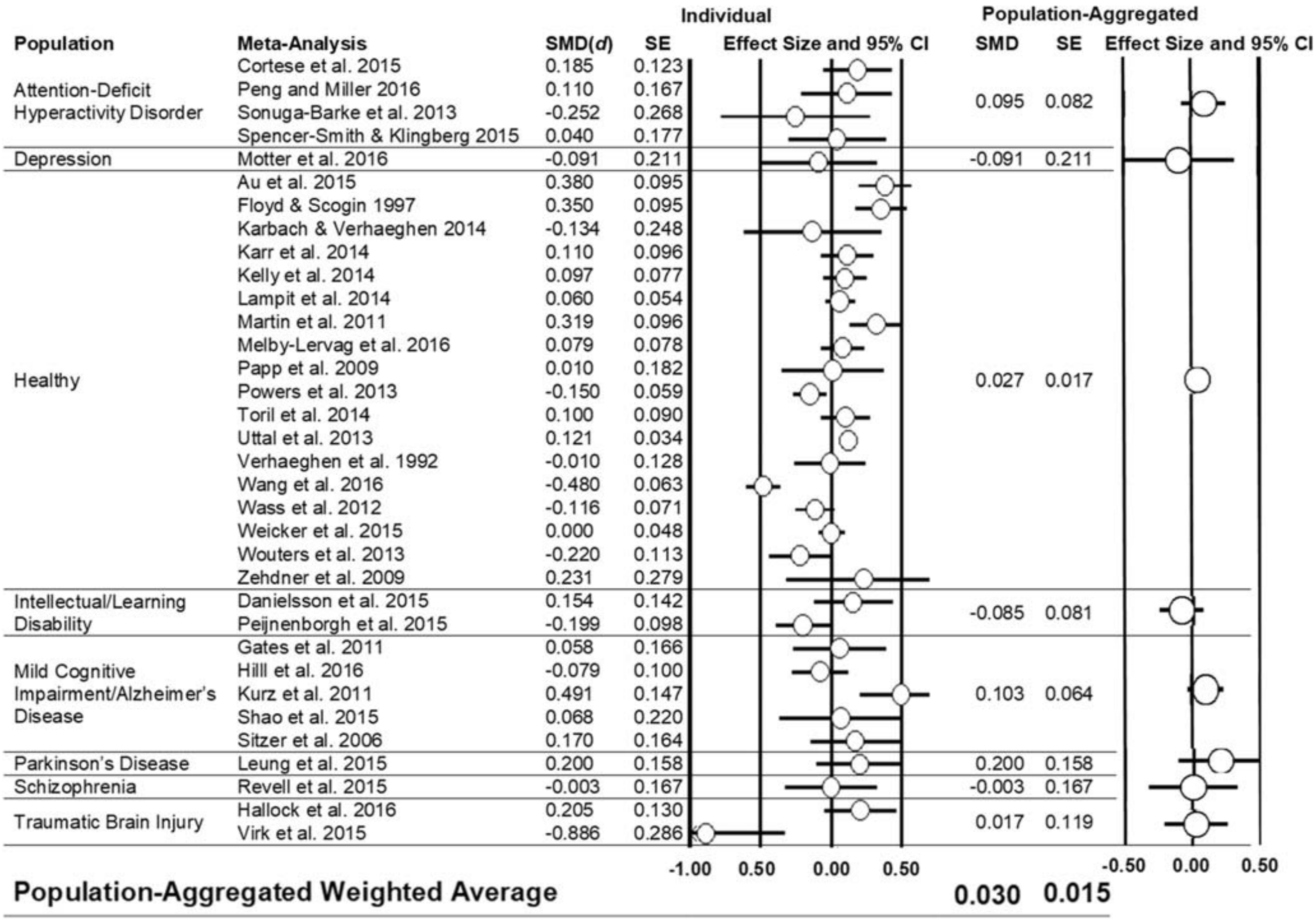
		Individual		Population-Aggregated	
Population	Meta-Analysis	SMD(*d*)	SE	SMD	SE
Attention-Deficit Hyperactivity Disorder	Cortese et al. 2015	0.436	0.135	0.343	0.082
	Peng and Miller 2016	0.250	0.148		
	Sonuga-Barke et al. 2013	0.301	0.351		
	Spencer-Smith & Klingberg 2015	0.330	0.158		
Depression	Motter et al. 2016	0.413	0.166	0.413	0.166
Healthy	Au et al. 2015	0.440	0.100	0.357	0.016
	Floyd & Scogin 1997	0.370	0.082		
	Karbach & Verhaeghen 2014	0.572	0.238		
	Karr et al. 2014	0.290	0.092		
	Kelly et al. 2014	0.340	0.063		
	Lampit et al. 2014	0.260	0.059		
	Martin et al. 2011	0.283	0.100		
	Melby-Lervag et al. 2016	0.295	0.085		
	Papp et al. 2009	0.180	0.181		
	Powers et al. 2013	0.320	0.046		
	Toril et al. 2014	0.370	0.090		
	Uttal et al. 2013	0.528	0.040		
	Verhaeghen et al. 1992	0.350	0.120		
	Wang et al. 2016	0.330	0.040		
	Wass et al. 2012	0.318	0.050		
	Weicker et al. 2015	0.370	0.048		
	Wouters et al. 2013	0.060	0.138		
	Zehdner et al. 2009	0.429	0.228		
Intellectual/Learning Disability	Danielsson et al. 2015	0.356	0.267	0.367	0.077
	Peijnenborgh et al. 2015	0.368	0.080		
Mild Cognitive Impairment/Alzheimer's Disease	Gates et al. 2011	0.109	0.184	0.354	0.065
	Hilll et al. 2016	0.282	0.124		
	Kurz et al. 2011	0.371	0.111		
	Shao et al. 2015	0.424	0.489		
	Sitzer et al. 2006	0.530	0.130		
Parkinson's Disease	Leung et al. 2015	0.353	0.174	0.353	0.174
Schizophrenia	Revell et al. 2015	0.099	0.166	0.099	0.166
Traumatic Brain Injury	Hallock et al. 2016	0.332	0.181	0.050	0.133
	Virk et al. 2015	-0.281	0.196		
Population-Aggregated Weighted Average				**0.344**	**0.023**

Fig. 3 Meta-meta-analysis of experimental vs. passive control comparisons. Across meta-analyses, cognitive training studies using passive control groups yield an overall effect size of $d = 0.344$. Positive effect sizes favor experimental groups

control comparison. If a study contained multiple intervention groups (whether an experimental or an active control group), all compared to the same control, then the raw scores of all intervention groups were averaged together first, and then compared to the control group in order to get one statistically independent net effect. However, if each intervention group was compared to its own respective control, then a Hedge's *g* effect size was calculated for each comparison and treated as independent (Borenstein 2009). There was never a situation in which multiple passive control groups were all compared to one intervention group. The end result was an independent set of effect sizes such that no condition within a study was represented more than once in the overall meta-analysis. The overall weighted average effect size comprises data from 396 objective cognitive assessments across all studies.

Risk of Bias and Heterogeneity

Per PRISMA guidelines, bias was assessed both within and between the studies included in the double-controlled meta-analysis. Potential bias within each study is described in Table S1, but no quantitative analysis was run due to the subjective nature of evaluating intra-study bias. Bias between studies was evaluated with a statistical analysis of publication bias, also referred to as the "file drawer problem." Publication bias refers to the phenomenon in which studies that report null results are less likely to be published. Therefore, the extant literature included in a meta-analysis is susceptible to a positive bias. We assessed publication bias qualitatively through the use of a funnel plot and quantitatively with Egger's regression for the double-controlled meta-analysis. The funnel plot is a graphical measure of publication bias or related small-study effects that plots effect sizes against standard errors. Under conditions of no bias, effect sizes should appear symmetric around the mean, with large studies (indexed by low standard errors) clustering tightly together near the top, but with increasing variability in effect size in smaller studies closer to the bottom. Under conditions of bias, where small or negative effect sizes are omitted from the literature, the plot will look more asymmetrical, especially with the small studies near the bottom which are more likely to be selected for positive or large effects. Egger's regression is a quantitative method of analyzing funnel plots and regresses the standard normal deviate, defined as the effect size divided by the standard error, against its precision (inverse standard error). With a perfectly symmetrical funnel plot, the intercept should be



Population	Meta-Analysis	SMD(*d*)	SE	SMD	SE
Attention-Deficit Hyperactivity Disorder	Cortese et al. 2015	0.185	0.123	0.095	0.082
	Peng and Miller 2016	0.110	0.167		
	Sonuga-Barke et al. 2013	-0.252	0.268		
	Spencer-Smith & Klingberg 2015	0.040	0.177		
Depression	Motter et al. 2016	-0.091	0.211	-0.091	0.211
Healthy	Au et al. 2015	0.380	0.095	0.027	0.017
	Floyd & Scogin 1997	0.350	0.095		
	Karbach & Verhaeghen 2014	-0.134	0.248		
	Karr et al. 2014	0.110	0.096		
	Kelly et al. 2014	0.097	0.077		
	Lampit et al. 2014	0.060	0.054		
	Martin et al. 2011	0.319	0.096		
	Melby-Lervag et al. 2016	0.079	0.078		
	Papp et al. 2009	0.010	0.182		
	Powers et al. 2013	-0.150	0.059		
	Toril et al. 2014	0.100	0.090		
	Uttal et al. 2013	0.121	0.034		
	Verhaeghen et al. 1992	-0.010	0.128		
	Wang et al. 2016	-0.480	0.063		
	Wass et al. 2012	-0.116	0.071		
	Weicker et al. 2015	0.000	0.048		
	Wouters et al. 2013	-0.220	0.113		
	Zehdner et al. 2009	0.231	0.279		
Intellectual/Learning Disability	Danielsson et al. 2015	0.154	0.142	-0.085	0.081
	Peijnenborgh et al. 2015	-0.199	0.098		
Mild Cognitive Impairment/Alzheimer's Disease	Gates et al. 2011	0.058	0.166	0.103	0.064
	Hilll et al. 2016	-0.079	0.100		
	Kurz et al. 2011	0.491	0.147		
	Shao et al. 2015	0.068	0.220		
	Sitzer et al. 2006	0.170	0.164		
Parkinson's Disease	Leung et al. 2015	0.200	0.158	0.200	0.158
Schizophrenia	Revell et al. 2015	-0.003	0.167	-0.003	0.167
Traumatic Brain Injury	Hallock et al. 2016	0.205	0.130	0.017	0.119
	Virk et al. 2015	-0.886	0.286		
Population-Aggregated Weighted Average				**0.030**	**0.015**

Fig. 4 Meta-meta-analysis of studies with active vs. passive controls. Across meta-analyses, cognitive training studies with passive controls yield an effect size that is $d = 0.030$ *larger* than studies with active controls. Positive effect sizes favor studies with passive controls, and suggest the possibility, but not the necessity, of placebo-like effects. However, the difference is only marginally significant ($p = 0.052$), and Bayesian statistics provide no support for the alternative hypothesis that any difference truly exists ($\mathrm{BF}_{10} = 0.859$)

close to zero, and the larger the deviation from zero, the greater the evidence for small-study effects such as publication bias. Negative values suggest bias in the direction of selecting larger effects in small studies. No assessment of bias was conducted on the meta-meta-analysis, given that meta-analyses do not seem to be systematically less likely to be published for reporting null results.

Heterogeneity was assessed using the I^2 statistic, which represents the percentage of total variation between studies that can be attributed to differences in true effect sizes rather than chance or sampling error alone. High I^2 values reflect greater heterogeneity and suggest that true differences exist between studies due to study design, population, or other factors other than sampling error alone. Conversely, a low I^2 value indicates homogeneity across studies and argues that the same basic effect is consistent across all studies, regardless of differences in study design, population, and other factors. Additionally, prediction intervals were calculated according to Borenstein et al. (2017) with the following formula: $d \pm t_{(\mathrm{df})}\sqrt{V_\mathrm{d} + \tau^2}$, where d is the mean effect size, t is the critical t value with a given degrees of freedom (df) equal to the number studies minus two, V_d is the variance of the effect size, and τ^2 is the variance of true effect sizes. The prediction interval is the range in which the true effect size would vary across 95% of heterogeneous populations/conditions.

Results

Meta-meta-analysis

The 34 effect sizes on the left-hand side of Fig. 2 represent the SMD between experimental and active control performance, as reported by each individual meta-analysis. The 8 effect sizes on the right represent the pooled SMD between experimental and active control performance within each population. Figure 3 displays the same information for the studies with passive controls. Aggregating the 8 population effect sizes together, the SMD among studies with active controls is $d = 0.308$ ($\mathrm{SE} = 0.020$, $p < 0.001$, $\mathrm{BF}_{10} = 1.35 \times 10^{23}$), whereas the SMD among studies with passive controls is $d = 0.344$ (SE = 0.023, $p < 0.001$, $BF_{10} = 1.83 \times 10^{26}$).

From both a frequentist and a Bayesian perspective, the analyses provide overwhelming support for a positive cognitive intervention effect with respect to both active and passive controls. However, in order to directly compare the relative difference between these effect sizes, we ran a paired within-meta-analysis comparison of the influence of active and passive controls (Fig. 4), revealing a very small, but nonetheless trending effect size difference of $d = 0.030$ (SE = 0.15, $p = 0.052$), numerically in favor of studies with passive controls outperforming studies with active controls. Despite the borderline significance of this finding, Bayesian analyses assessing the strength of evidence in favor of the alternative hypothesis find no support that there is any difference between either type of control group ($BF_{10} = 0.859$). Even the sensitivity analyses revealed little support for very small ($r = 0.01$, $BF_{10} = 1.093$), small ($r = 0.1$, $BF_{10} = 1.572$), or moderate ($r = 0.3$, $BF_{10} = 1.606$) effects.

Double-Controlled Meta-analysis

Figures 5 and 6 show significant cognitive intervention effects against both active ($g = 0.250$, SE = 0.045, $p < 0.001$, $BF_{10} = 1.138 \times 10^5$) and passive controls ($g = 0.309$, SE = 0.046, $p < 0.001$, $BF_{10} = 1.832 \times 10^8$), with Bayesian analyses providing decisive evidence for the alternative hypothesis in both cases. However, when directly comparing the performance of active and passive controls to each other (Fig. 7), no significant performance differences were found ($g = 0.058$, SE = 0.044, $p = 0.194$) and in fact, Bayesian analyses strongly supported this null finding ($BF_{01} = 12.046$). However, sensitivity analyses on the Bayes factor showed weak evidence for the null (but no evidence for the alternative) when aiming to detect very small ($r = 0.01$; $BF_{01} = 1.017$) or small ($r = 0.1$; $BF_{01} = 1.660$) effects and showed substantial evidence when allowing for more moderate effects ($r = 0.3$; $BF_{01} = 3.78$).

Publication Bias

In the double-controlled meta-analysis, the likelihood for publication bias is small because the studies in our sample are generally published based on the merits of the experimental groups and not on the performance of the control groups. Nevertheless, we cannot exclude the possibility that publication bias may be selecting more strongly for the experimental/active control comparison rather than the experimental/passive control comparison. To examine this, we carried out an analysis of publication bias for both comparisons using funnel plots (Fig. 8) and Egger's regression intercept (Egger et al. 1997). In the experimental/active control comparison, Egger's intercept was − 0.447 (SE = 0.425, $p = 0.299$) and with the experimental/passive control comparison, Egger's intercept was − 0.163 (SE = 0.517, $p = 0.754$). Neither analysis reached significance to endorse small-study effects, despite reasonable meta-analytic power with 44 effect sizes. More critically for our analyses, however, there is little evidence that bias, even if it exists, systematically affects comparisons with active control groups differently than comparisons with passive control groups, as their confidence intervals are highly overlapping.

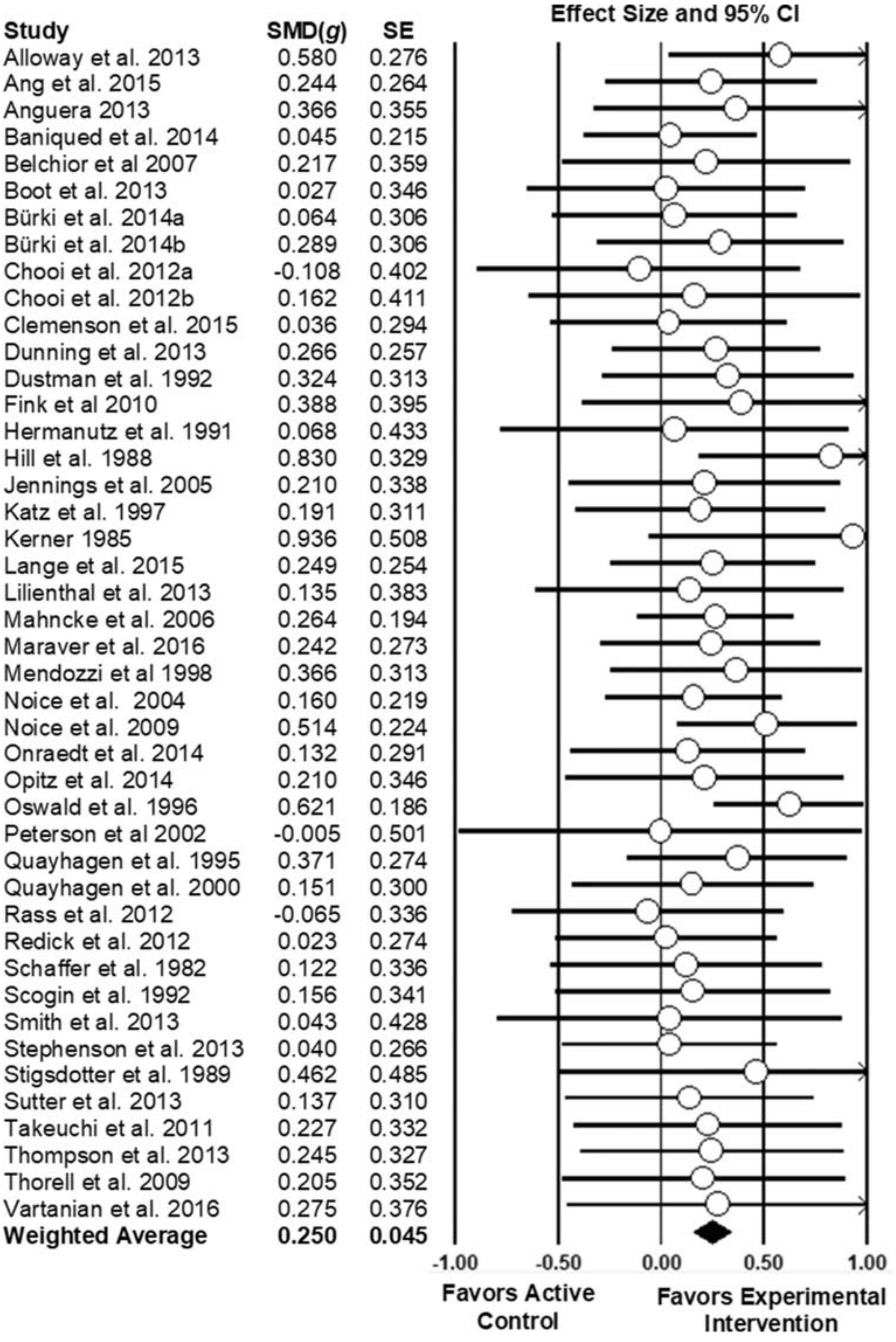
Heterogeneity

Heterogeneity was assessed in the meta-meta-analysis using the full sample of 34 meta-analyses, rather than the 8 collapsed population groups in order to prevent the averaging over of heterogeneity that may exist within populations. In the comparison with active controls, significant heterogeneity was found ($Q = 115.166$, $I^2 = 71.346$, $p < 0.001$, $\tau^2 = 0.022$), with the 95% prediction interval suggesting that 95% of true effects range from $d = 0.003$ to $d = 0.613$ (Borenstein et al. 2017). In the comparison with passive controls, no significant heterogeneity was found ($Q = 7.789$, $I^2 = 10.130$, $p = 0.352$, $\tau^2 = 0.001$), with 95% of true effects ranging from $d = 0.264$ to $d = 0.424$. In both cases, heterogeneity estimates likely represent a lower bound since many meta-analyses shared primary studies with each other, thereby potentially masking some heterogeneity effects.

Within the double-controlled meta-analysis, no evidence for heterogeneity was found ($Q = 8.225$, $I^2 = 0.000$, $p = 1.000$, $\tau^2 = 0.000$), with 95% of true effects ranging from $g = -0.032$ to $g = 0.148$.

Moderator Analyses

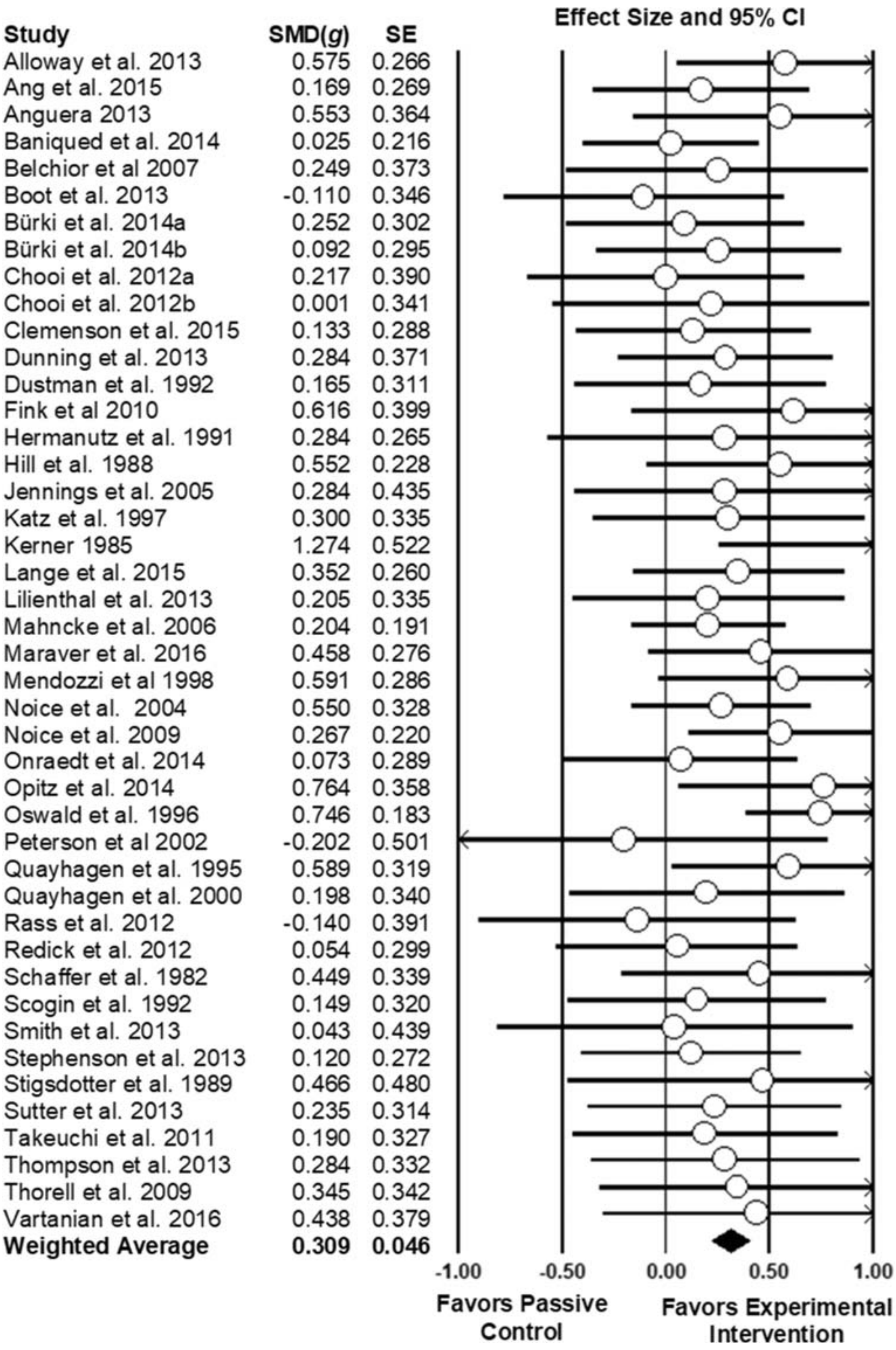
Despite the lack of a significant main effect and the lack of heterogeneity in the double-controlled meta-analysis, we nevertheless attempted an exploratory moderator analysis in order to reveal whether there might be indications for differential placebo effects as a function of a specific subset of our data, and if so, whether there were specific populations or types of outcome measures or types of active control designs that might be particularly prone to placebo effects. None of the population effects approached significance: clinical populations ($g = 0.090$, SE = 0.107, $p = 0.400$, $BF_{01} = 6.606$), healthy participants ($g = 0.051$, SE = 0.049, $p = 0.296$, $BF_{01} = 14.811$), younger participants under 60 years old ($g = 0.055$, SE = 0.065, $p = 0.393$, $BF_{01} = 13.419$), older participants over 60 years old ($g = 0.060$, SE = 0.061, $p = 0.326$, $BF_{01} = 12.581$). None of the outcome measures approached significance: visual outcome measures ($g = 0.056$, SE = 0.056, $p = 0.320$, $BF_{01} = 13.408$), verbal outcome measures ($g = 0.040$, SE = 0.049, $p = 0.417$, $BF_{01} = 18.447$), mixed modality outcome measures ($g = 0.078$, SE = 0.056, $p = 0.167$, $BF_{01} = 8.603$), fluid intelligence outcomes ($g = 0.023$, SE = 0.068,



Study	SMD(g)	SE
Alloway et al. 2013	0.580	0.276
Ang et al. 2015	0.244	0.264
Anguera 2013	0.366	0.355
Baniqued et al. 2014	0.045	0.215
Belchior et al 2007	0.217	0.359
Boot et al. 2013	0.027	0.346
Bürki et al. 2014a	0.064	0.306
Bürki et al. 2014b	0.289	0.306
Chooi et al. 2012a	-0.108	0.402
Chooi et al. 2012b	0.162	0.411
Clemenson et al. 2015	0.036	0.294
Dunning et al. 2013	0.266	0.257
Dustman et al. 1992	0.324	0.313
Fink et al 2010	0.388	0.395
Hermanutz et al. 1991	0.068	0.433
Hill et al. 1988	0.830	0.329
Jennings et al. 2005	0.210	0.338
Katz et al. 1997	0.191	0.311
Kerner 1985	0.936	0.508
Lange et al. 2015	0.249	0.254
Lilienthal et al. 2013	0.135	0.383
Mahncke et al. 2006	0.264	0.194
Maraver et al. 2016	0.242	0.273
Mendozzi et al 1998	0.366	0.313
Noice et al. 2004	0.160	0.219
Noice et al. 2009	0.514	0.224
Onraedt et al. 2014	0.132	0.291
Opitz et al. 2014	0.210	0.346
Oswald et al. 1996	0.621	0.186
Peterson et al 2002	-0.005	0.501
Quayhagen et al. 1995	0.371	0.274
Quayhagen et al. 2000	0.151	0.300
Rass et al. 2012	-0.065	0.336
Redick et al. 2012	0.023	0.274
Schaffer et al. 1982	0.122	0.336
Scogin et al. 1992	0.156	0.341
Smith et al. 2013	0.043	0.428
Stephenson et al. 2013	0.040	0.266
Stigsdotter et al. 1989	0.462	0.485
Sutter et al. 2013	0.137	0.310
Takeuchi et al. 2011	0.227	0.332
Thompson et al. 2013	0.245	0.327
Thorell et al. 2009	0.205	0.352
Vartanian et al. 2016	0.275	0.376
Weighted Average	**0.250**	**0.045**

Fig. 5 Meta-analysis of experimental/active control comparisons. Within our sample of double-controlled studies, the effect size of cognitive training on objective cognitive tests when compared to active controls is $g = 0.250$. Outcomes that were specifically trained were excluded from analysis; thus, this effect size only reflects transfer to untrained tasks

$p = 0.731$, $\mathrm{BF}_{01} = 17.167$), working memory outcomes ($g = 0.049$, $\mathrm{SE} = 0.054$, $p = 0.361$, $\mathrm{BF}_{01} = 15.290$), process-based outcomes (including all cognitive laboratory tasks; $g = 0.058$, $\mathrm{SE} = 0.047$, $p = 0.220$, $\mathrm{BF}_{01} = 12.493$), and non-process-based outcomes (including crystallized intelligence and motor tasks; $g = -0.128$, $\mathrm{SE} = 0.098$, $p = 0.189$, $\mathrm{BF}_{01} = 13.313$). None of the active control design effects approached significance: working memory interventions ($g = 0.054$, $\mathrm{SE} = 0.104$, $p = 0.604$, $\mathrm{BF}_{01} = 10.654$), memory-based interventions ($g = 0.056$, $\mathrm{SE} = 0.100$, $p = 0.572$, $\mathrm{BF}_{01} = 10.777$), process-based interventions ($g = 0.071$, $\mathrm{SE} = 0.07$, $p = 0.311$, $\mathrm{BF}_{01} = 10.723$), non-process-based interventions (including socio-emotional stimulation, reading books or watching movies, motor tasks, etc.; $g = 0.049$, $\mathrm{SE} = 0.057$, $p = 0.396$, $\mathrm{BF}_{01} = 15.071$), attention-based interventions ($g = 0.082$, $\mathrm{SE} = 0.171$, $p = 0.634$, $\mathrm{BF}_{01} = 6.649$), or speed-based interventions



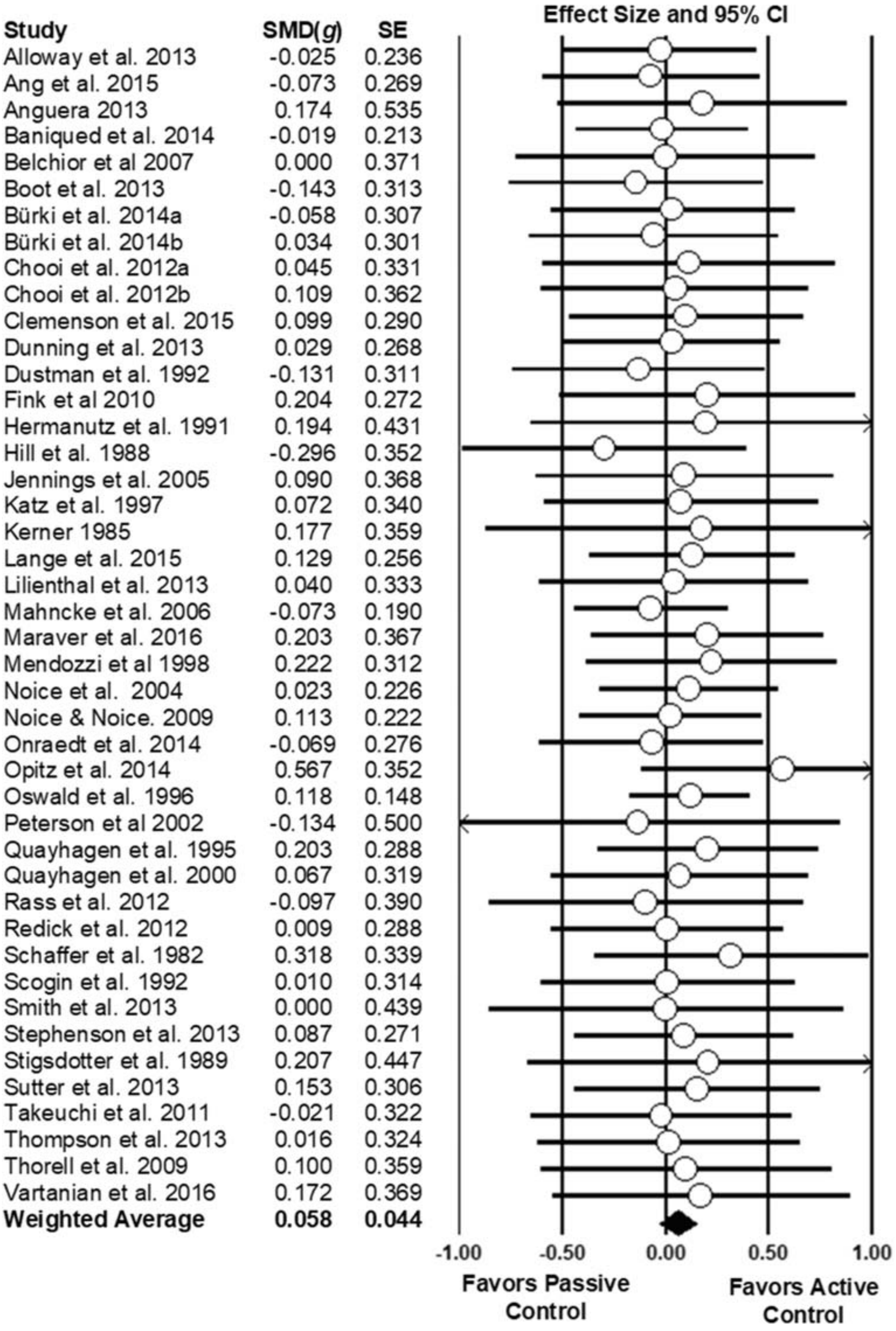
Study	SMD(*g*)	SE
Alloway et al. 2013	0.575	0.266
Ang et al. 2015	0.169	0.269
Anguera 2013	0.553	0.364
Baniqued et al. 2014	0.025	0.216
Belchior et al 2007	0.249	0.373
Boot et al. 2013	-0.110	0.346
Bürki et al. 2014a	0.252	0.302
Bürki et al. 2014b	0.092	0.295
Chooi et al. 2012a	0.217	0.390
Chooi et al. 2012b	0.001	0.341
Clemenson et al. 2015	0.133	0.288
Dunning et al. 2013	0.284	0.371
Dustman et al. 1992	0.165	0.311
Fink et al 2010	0.616	0.399
Hermanutz et al. 1991	0.284	0.265
Hill et al. 1988	0.552	0.228
Jennings et al. 2005	0.284	0.435
Katz et al. 1997	0.300	0.335
Kerner 1985	1.274	0.522
Lange et al. 2015	0.352	0.260
Lilienthal et al. 2013	0.205	0.335
Mahncke et al. 2006	0.204	0.191
Maraver et al. 2016	0.458	0.276
Mendozzi et al 1998	0.591	0.286
Noice et al. 2004	0.550	0.328
Noice et al. 2009	0.267	0.220
Onraedt et al. 2014	0.073	0.289
Opitz et al. 2014	0.764	0.358
Oswald et al. 1996	0.746	0.183
Peterson et al 2002	-0.202	0.501
Quayhagen et al. 1995	0.589	0.319
Quayhagen et al. 2000	0.198	0.340
Rass et al. 2012	-0.140	0.391
Redick et al. 2012	0.054	0.299
Schaffer et al. 1982	0.449	0.339
Scogin et al. 1992	0.149	0.320
Smith et al. 2013	0.043	0.439
Stephenson et al. 2013	0.120	0.272
Stigsdotter et al. 1989	0.466	0.480
Sutter et al. 2013	0.235	0.314
Takeuchi et al. 2011	0.190	0.327
Thompson et al. 2013	0.284	0.332
Thorell et al. 2009	0.345	0.342
Vartanian et al. 2016	0.438	0.379
Weighted Average	**0.309**	**0.046**

Fig. 6 Meta-analysis of experimental/passive control comparisons. Within our sample of double-controlled studies, the effect size of cognitive training on objective cognitive tests when compared to active controls is $g = 0.309$. Outcomes that were specifically trained were excluded from analysis; thus, this effect size only reflects transfer to untrained tasks

($g = 0.129$, $\mathrm{SE} = 0.178$, $p = 0.572$, $\mathrm{BF}_{01} = 5.587$). See Table S2 for examples of each moderator category.

Discussion

The goal of the present work was to determine potential performance differences between active and passive control groups in cognitive training studies (e.g., Melby-Lervåg et al. 2016; Shipstead et al. 2012). Overall, our results do not provide evidence within a broad spectrum of the existing



Study	SMD(*g*)	SE
Alloway et al. 2013	-0.025	0.236
Ang et al. 2015	-0.073	0.269
Anguera 2013	0.174	0.535
Baniqued et al. 2014	-0.019	0.213
Belchior et al 2007	0.000	0.371
Boot et al. 2013	-0.143	0.313
Bürki et al. 2014a	-0.058	0.307
Bürki et al. 2014b	0.034	0.301
Chooi et al. 2012a	0.045	0.331
Chooi et al. 2012b	0.109	0.362
Clemenson et al. 2015	0.099	0.290
Dunning et al. 2013	0.029	0.268
Dustman et al. 1992	-0.131	0.311
Fink et al 2010	0.204	0.272
Hermanutz et al. 1991	0.194	0.431
Hill et al. 1988	-0.296	0.352
Jennings et al. 2005	0.090	0.368
Katz et al. 1997	0.072	0.340
Kerner 1985	0.177	0.359
Lange et al. 2015	0.129	0.256
Lilienthal et al. 2013	0.040	0.333
Mahncke et al. 2006	-0.073	0.190
Maraver et al. 2016	0.203	0.367
Mendozzi et al 1998	0.222	0.312
Noice et al. 2004	0.023	0.226
Noice & Noice. 2009	0.113	0.222
Onraedt et al. 2014	-0.069	0.276
Opitz et al. 2014	0.567	0.352
Oswald et al. 1996	0.118	0.148
Peterson et al 2002	-0.134	0.500
Quayhagen et al. 1995	0.203	0.288
Quayhagen et al. 2000	0.067	0.319
Rass et al. 2012	-0.097	0.390
Redick et al. 2012	0.009	0.288
Schaffer et al. 1982	0.318	0.339
Scogin et al. 1992	0.010	0.314
Smith et al. 2013	0.000	0.439
Stephenson et al. 2013	0.087	0.271
Stigsdotter et al. 1989	0.207	0.447
Sutter et al. 2013	0.153	0.306
Takeuchi et al. 2011	-0.021	0.322
Thompson et al. 2013	0.016	0.324
Thorell et al. 2009	0.100	0.359
Vartanian et al. 2016	0.172	0.369
Weighted Average	**0.058**	**0.044**

Fig. 7 Meta-analysis of active/passive control comparisons. Within our sample of double-controlled studies, the within-study performance difference between active and passive control groups is not significant (g = 0.058), and Bayesian statistics support the null hypothesis ($\text{BF}_{01} = 12.046$). Outcomes that were specifically trained were excluded from analysis

cognitive intervention literature that the type of control group used in a study pervasively influences results on objective neuropsychological or ability measures. Two complementary approaches led us to this conclusion.

First, a meta-meta-analysis consisting of 34 individual meta-analyses revealed significant intervention effects in studies using both types of control, which refutes the notion that cognitive training effects get erased when active controls are used (c.f., Melby-Lervåg et al. 2016). Moreover, the small difference between the two effect sizes ($d = 0.03$; Fig. 4) was found to be only marginally significant ($p = 0.052$) in favor of stronger effects among studies using only passive controls. Additionally, Bayesian analyses offer very little evidence to support the existence of true differences, with Bayes factors ranging from a negligible 0.859 to 1.606 using a wide range of plausible priors. Even disregarding the support for the null, the

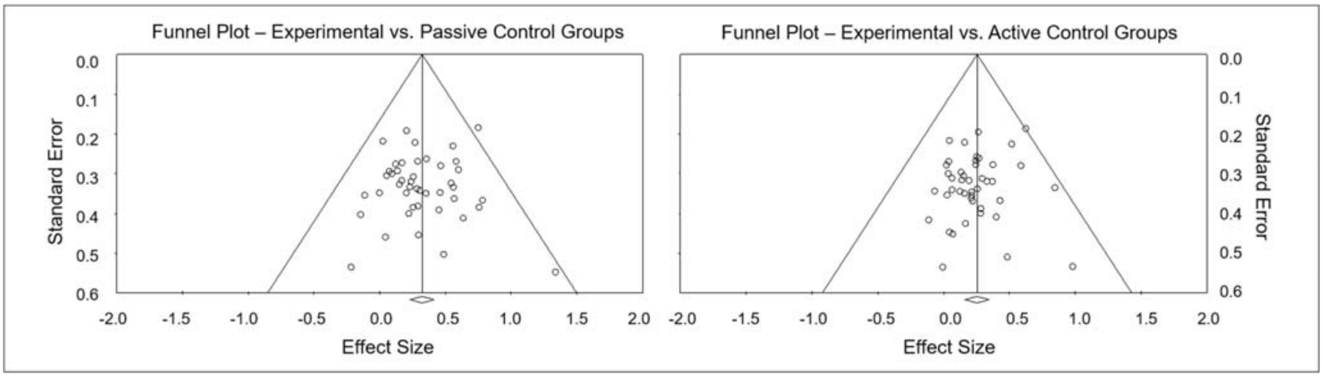


Fig. 8 Funnel plots from comparisons of experimental with passive and active control groups (double-controlled studies). No asymmetry was statistically detectable in the funnel plots, neither for the comparison of experimental groups with passive controls (left) nor the comparison with active controls (right). More critically, the degree of asymmetry, though non-significant, is similar between both comparisons, suggesting that if bias does exist, it does not systematically affect one type of control group over the other

effect estimate of $d = 0.03$ seems of little practical relevance, especially given that the intervention effects of the experimental groups were found to be 0.308 and 0.344 (Figs. 2 and 3), an over tenfold difference.

Although these meta-meta-analytic results support the notion that experimental results from studies with passive or active controls do not differ, interpretation is still limited as these results are correlational in nature and there could be a variety of other factors associated with the control group design choice that may also moderate the effect size (see Au et al. 2016). In order to provide further evidence for or against an effect of control group type, we conducted an additional meta-analysis on cognitive training studies that used both a passive and an active control group *within* the same experiment. In this way, any other correlated factors that may exist within the same study are controlled for.

Similar to our first analysis, we found no significant difference in performance between passive and active controls when compared directly ($g = 0.058$, $p = 0.194$). Despite a numerical advantage of active controls, once again Bayesian statistics provide no evidence for any true effects. In fact, the null model is 12 times more likely than the alternative, suggesting strong evidence for the equivalence of active and passive control group performance. Our sensitivity analyses show that selecting a moderate prior of $r = 0.3$, which places 50% of the prior probability mass on effect sizes ranging from -0.3 to $+0.3$, there is still substantial evidence for the null ($BF_{01} = 3.798$). Even when using liberal priors of $r = .01$ or $r = 0.1$ to capture very small or small effects, Bayes factors for the null hypothesis are 1.017 and 1.660, respectively, which contain no to little evidential value one way or the other. Thus, across a range of analytical approaches, there is accumulated evidence that the use of passive or active control groups in the current cognitive training literature *cannot* explain moderate to large effects on objective outcome measures, and there is no indication one way or another that even small effects exist. In the following sections, we explore whether these effects may still exist in a subsection of the data, discuss possible reasons why they are absent, and finally offer ideas for next steps and future directions.

Do Some Studies Show More Control Group Differences than Others?

Having established the absence of an overall effect of control type within the cognitive training literature at large, we next sought to determine whether certain study or active control design choices influence the ability to detect these effects. At the meta-meta-analytic level, we detected significant heterogeneity among studies with active controls, but not among studies with passive controls. This suggests the possibility that not all active controls are created equal, and some may provide more rigorous controls than others, thus erasing experimental effects in those studies. For instance, the prediction interval around studies with active controls extends down to $d =$ 0.003, suggesting that a subset of studies show zero true effects, whereas the effect sizes from studies with passive controls all hover fairly homogenously around the mean estimate of $d = 0.344$. However, before endorsing the conclusion that some active controls are designed more rigorously than others, we first reiterate that these analyses are correlational, and alternative explanations still exist. For example, we have previously analyzed this precise pattern of effects in a specific subset of studies examining the influence of n-back training on fluid intelligence measures, and there we demonstrated that this pattern was *not* driven by any differences in control group performance (Au et al. 2015, 2016). Rather, we observed a curious and as yet unexplained underperformance of experimental groups in the studies that happened to use active controls, leading to the smaller effect sizes observed in these studies relative to those that used passive controls (Au et al. 2015, 2016). Moreover, Fig. 4 in our current data, which summarizes the effect size advantage of studies with passive controls over those with active controls, shows extreme effects in both directions, with the prediction interval ranging from $d = -0.347$ to $d = 0.407$, suggesting the existence of a subset of studies that favor active control performance as well as a subset of studies that favor passive control performance. Thus, it is difficult to convincingly argue that our observed meta-meta-analytic heterogeneity is driven by any systematic advantage of studies that employ one control type over another.

When analyzing heterogeneity in our double-controlled meta-analysis, which better approximates a causal framework by controlling for any within-study idiosyncrasies, we detected no significant heterogeneity in any of the three comparisons, including the direct active/passive control comparison. However, the lack of heterogeneity may be at least in part attributed to the small sample sizes used in cognitive training studies, leading to wide, overlapping confidence intervals that can potentially mask the existence of true heterogeneity. Notably, even the experimental/control comparisons did not demonstrate heterogeneity, despite using a wide range of intervention tasks and populations. Thus, to further probe possible heterogeneity, additional moderator analyses were run attempting to detect any possible influences of outcome measures, population, or active control design. However, none of the analyses revealed any differences (Table S2). Thus, the cognitive training effect size is fairly homogenous within this dataset of 42 studies, and no argument can be convincingly made that a systematic advantage for active controls exists in any subset of studies.

Why Is There No Difference Between Passive and Active Control Groups?

We see at least two possible conclusions that can be drawn from our data. First, some may infer that the lack of difference between passive and active controls indicates that active controls as they have been used in the extant cognitive training literature are insufficiently designed to be able to reliably capture such differences. Second, it is also possible instead that the lack of difference indicates that placebo and other non-specific artifacts do not systematically and pervasively occur with objective outcomes in the cognitive training literature. Neither conclusion can be ruled out with the current data and both might be true to some extent. In the following, we elaborate more on both possible interpretations.

Interpretation 1: Current Active Control Groups Are Insufficiently Designed To Elicit Placebo Effects

Within our sample of double-controlled studies, active control designs ranged wildly in terms of the degree to which they approximated the experimental task and the degree to which they might be considered a "believable" intervention by participants. Designs that are too dissimilar from their experimental counterparts may not fully control for all possible confounds, while designs that are overly similar risk inadvertently controlling out relevant training effects. For example, on the more dissimilar end, one study in our meta-analytic sample examined the effects of computerized memory and attention training on schizophrenic patients and had control group participants simply watch television for the same amount of time (Rass et al. 2012). While this "active" control may control for non-specific intervention effects related to experimenter contact and time spent on a task, other potentially influential factors differ between groups such as perceived cognitive effort and expectations of improvement. However, opposite problems can arise as well on the other end of the spectrum when active control interventions overly resemble their experimental counterparts. Take for instance Opitz et al. (2014), who compared visual n-back training (experimental intervention) to auditory n-back training (active control) to improve Chinese vocabulary learning. Although in some respects this seems like an ideal active control because it effectively isolates a single hypothesized factor (processing of visual stimuli in working memory) while keeping all other intervention characteristics identical, interpretation must proceed carefully in a field like working memory training where the underlying mechanisms of positive training effects are still not well understood. For example, it is unclear to what extent positive training results are modality-specific (e.g., Jaeggi et al. 2014; Schneiders et al. 2011), and therefore whether both auditory and visual working memory training might train similar and more general underlying processes. It is very difficult to design tasks that rely purely on one modality as researchers cannot control any cross-modality strategies participants choose to use (e.g., verbal encoding of visual information). Furthermore, irrespective of that, working memory and related processes are also known to involve both modality-specific as well as modality-general functional networks (Hsu et al. 2017; Li et al. 2014), and thus, researchers cannot rule out that modality-general improvements may arise from ostensibly modality-specific training that also benefits performance on transfer measures. Indeed, the active/passive control effect size in Opitz et al. (2014) is the largest in our meta-analytic sample ($g = 0.582$), and it is unknown whether this came about as a result of the auditory working memory intervention producing real training gains, or by the induction of non-specific placebo-like effects that inflated performance at posttest over and above the passive control group, or by some combination of the two.

As illustrated, proper design of active control interventions is not a simple matter, and the border between what constitutes a control task and an experimental task is fuzzy (Rebok 2015). In fact, there were several instances in our meta-analytic sample in which researchers chose a task as an active control in hopes of creating a believable intervention that elicits no meaningful cognitive benefit (Boot et al. 2008; Opitz et al. 2014; Stephenson and Halpern 2013; Thompson et al. 2013; Vartanian et al. 2016), while other researchers actually decided to use those very same tasks as experimental interventions to elicit cognitive improvement (Jaeggi et al. 2008; Smith et al. 2013; Thorell et al. 2009; Vartanian et al. 2016).

Interpretation 2: Placebo and Other Non-specific Artifacts Do Not Pervasively Occur with Objective Cognitive Measures

Although the existence of placebo and other non-specific intervention effects is not controversial, it should not be universally assumed that these effects are pervasive and can be measured reliably across all domains. The concept of a placebo effect first originated in the medical field where the primary dependent variable is the (often subjectively assessed) symptomatology improvement of the patient, such as the perception of pain (Beecher 1955). However, evidence for placebo-like enhancement on objective neuropsychological outcomes, such as the ones used in the current meta-analysis, has been more difficult to elicit (Green et al. 2001; Hróbjartsson and Gøtzsche 2001, 2004, 2010; Looby and Earleywine 2011; Schwarz and Büchel 2015). For example, in a recent working memory-based intervention study, Tsai et al. (2018) induced positive and negative participant expectations but did not observe any group differences (experimental vs. active control) in objective performance as a function of induction. Instead, only the experimental group demonstrated transfer to an untrained working memory task, irrespective of having negative expectations, while the control group showed no improvement despite having positive expectations. Similarly, Schwarz and Büchel (2015) induced expectations of cognitive improvement in participants during an inhibitory control task, and although participants did indeed believe they performed better, their objective performance itself did not change. Green et al. (2001) likewise found no consistent improvements on a battery of executive function tests after placebo glucose administration, observing improvement on only one test, out of eight. Furthermore, a meta-analysis of 27 clinical trials across a diverse set of health conditions compared active control groups to no-treatment groups and found no overall placebo advantage on objective measures of symptom improvement despite demonstrating self-reported improvements on subjective outcomes (Hróbjartsson and Gøtzsche 2001). Nevertheless, isolated incidents of placebo improvement on objective neuropsychological measures have also been documented (Foroughi et al. 2016). The rationale for subjective improvements after placebo induction is clear, since they operate in the domain of beliefs and expectations, leading to response biases. However, the pathway to objective improvements is more indirect since it relies on the power of belief to exert some physiological change in the body and for that change to become relevant to the outcome being measured. While such physiological changes have certainly been documented, such as changes in neurotransmission and opioid receptor activity during placebo-induced analgesia or changes in brain glucose metabolism with placebo anti-depressants, (Benedetti et al. 2005; Price et al. 2008; Wager and Atlas 2015), these objective effects are less well understood and less consistent outside the pain and analgesia literature (Benedetti et al. 2005; Hróbjartsson and Gøtzsche 2001). Moreover, it is not always clear which types of objective outcomes are affected by these physiological changes and under what conditions, so it should not be assumed by default that cognitive improvements automatically fall under this umbrella. In fact, our current data, in accordance with the literature, suggest that if these cognitive effects exist, they are not easy to induce even when studies explicitly aim to do so.

How To Move Forward

What could be the resolution to this conundrum? First, it seems that the research community needs to recognize the problem and direct more efforts into specifying and quantifying any non-specific intervention effects that may exist, such as placebo effects, experimenter demand characteristics, and other influences. It is insufficient to simply rely on active control designs to rule out these influences. Instead, data are required to measure the extent to which these factors influence performance of both experimental and control groups, for different types of interventions and different types of outcome measures. For instance, it would be beneficial if studies would routinely assess (and/or manipulate) expectations, motivation, fatigue, and other psychological phenomena, a practice which is rarely done in the current literature. A few enterprising studies have already taken this route (Foroughi et al. 2016; Katz et al. 2018; Tsai et al. 2018). However, to date, they have yielded inconsistent evidence, showing that expectations may (Foroughi et al. 2016) or may not (Tsai et al. 2018) influence transfer results from cognitive training and that different motivational influences affect cognitive performance on different tasks in different ways (Cerasoli et al. 2014; Katz et al. 2018). Although the complex pattern of results may be daunting to tackle, it is imperative that we continue to measure these effects in cognitive training studies in order to develop a better understanding of their influence and the conditions under which they manifest (or not). Furthermore, long-term follow-ups should be incorporated whenever possible, as any confounding effects of motivation and fatigue are more likely to wash out after a period of time in order to allow a more pure measurement of training effects (e.g., Klauer and Phye 2008). Additionally, a stronger emphasis on the underlying mechanisms of cognitive training would be fruitful in elucidating the extent to which an active control task may overlap with an experimental task without accidentally incurring meaningful benefits. This line of research is already underway (Buschkuehl et al. 2014; Dahlin et al. 2008; Hsu et al. 2013; Hussey et al. 2017; Jaeggi et al. 2014; Katz et al. 2018; Salmi et al. 2018) but we are still far from a satisfactory understanding of the mechanisms under which transfer of cognitive training occurs.

Until our understanding of these issues reach maturity, it is difficult to strategically and rigorously design appropriate active controls, and we caution against an exclusive reliance on studies using active controls, as has become the current trend (i.e., Melby-Lervåg and Hulme 2013; Melby-Lervåg et al. 2016). Rather, interpretations about the efficacy of cognitive training should be based on the totality of the extant literature, and it would be imprudent to dismiss a meaningful portion of that literature on the assumption that passive controls perform differently than active controls. Moreover, there are advantages to the use of passive controls that are often overlooked: Passive control groups provide a consistent and generally reliable control for retest effects across studies. Therefore, effect sizes derived from passive controls can be compared to the same standard across studies for different experimental interventions. This is supported by our meta-meta-analytic observation of fairly homogenous effect sizes across the spectrum of studies using passive controls, but not among those using active controls. Furthermore, they are more cost-effective and easier to implement, which may be useful for preliminary phase 1 trials, or small proof-of-principle studies (C. S. Green et al. 2019; Willis 2001). Reducing these impediments has several benefits. First, it allows for a relatively low-investment opportunity to evaluate the feasibility of new interventions, as it curtails the expenses associated with paying research staff and participants to be involved with the design and the implementation of an active control intervention. Second, it allows for all basic efficacy intervention studies to be easily compared on a meta-analytic level because they would all have a homogenous control that purely accounts for retest effects. Although interpretation would have to proceed with the knowledge that non-specific confounds may exist (a problem which currently has no convincing solution even with the use of active controls), at the very least, interventions could be compared to each other in order to determine relative effects.

To be clear, we are not dismissing the use of active controls simply because they have failed to outperform passive controls in the past. Rather, we recommend a more nuanced approach in using and interpreting intervention data that rely on both passive and active controls (see also C. S. Green et al. 2019). Until future research can convincingly elucidate both the nature and extent of placebo and placebo-like effects on objective cognitive outcomes and more clearly delineate the boundary between a control task that effectively induces these non-specific effects versus a control task that inadvertently trains relevant cognitive processes, a sensible role for the use of passive controls may continue to exist. We suggest active controls, on the other hand, can be used as a second tier strategy to test interventions that have at least passed the basic efficacy phase with passive controls. Here, researchers can strategically design the active control to assess and rule out specific and quantifiable placebo-like effects such as expectancies or Hawthorne effects. Additionally, the active control can focus on isolating properties of the training task in order to get at the candidate mechanisms that may underlie intervention efficacy (e.g., Hussey et al. 2017; Oelhafen et al. 2013). In an ideal-world scenario, passive and active controls should both be used within the same study to more rigorously test for the existence and extent of placebo-like effects.

Limitations

In our attempt to garner a comprehensive and far-reaching perspective from the literature on this issue of passive and active control group comparisons, we necessarily invite some limitations to our data, chief of which is the noise inherent in dealing with such a large and diverse dataset. For example, the definition of what an active control entails differs between studies as there is no current consensus or gold standard. Although we endeavored to reduce our own subjectivity by yielding to the definitions and categorizations of individual authors, this variability also potentially renders active and passive controls more similar to each other than if one universal standard was applied to all active controls. Although the heterogeneous nature of active controls across studies is a limitation we cannot get around, we nevertheless point out that our moderator analyses demonstrated that all categories of active controls that we analyzed performed similarly to passive controls at the meta-analytic level.

We must also acknowledge the considerable heterogeneity in how individual meta-analyses reported their effect sizes, some using Cohen's *d*, while others used Hedges' *g*, and some calculating differences in gain scores between experimental and control groups, while others used only posttest scores, and the different types of outcomes each meta-analysis accepted into their analysis. However, Cohen's *d* and Hedges' *g* are almost identical to each other except when sample sizes are very small (e.g., < 10), and very few cognitive training studies have sample sizes in the single digits. Additionally, with the randomized designs of most cognitive training studies, effect sizes calculated from only posttest data or taking into account pretest as well generally perform fairly similarly and tend to agree closely with each other, especially when averaging across many studies (Au et al. 2016). Furthermore, when comparing studies with passive and active controls to each other in the meta-meta-analysis, we were careful to conduct this analysis in a within-meta-analysis manner so that these idiosyncrasies are controlled out. Therefore, we argue that the effect sizes between meta-analyses are comparable nonetheless and, moreover, point out that any differences that may exist occur randomly and non-systematically.

Although our double-controlled meta-analysis circumvents some of the noisiness inherent in the meta-meta-analysis by controlling for within-study differences, it is also subject to its own limitations in that it may represent a unique subset of studies whose generalizability to the rest of the field is not fully certain. For example, studies that employ both passive and active control groups may be more rigorous or may be better funded. They also tend to be more recent, as the average year of publication within our meta-analysis was 2009. Therefore, any interpretations of this meta-analytic dataset must be made with these potential confounds and biases in mind. However, it is heartening to see that the results converge with our broader meta-meta-analysis as well, and we encourage readers to consider both analyses together when interpreting our data as they both have their own unique strengths and weaknesses that complement each other.

Conclusions

Our two complementary and comprehensive meta-analyses, which aggregate data from a very substantial portion of the cognitive training literature to date (1524 studies), demonstrate no evidence that control group type meaningfully influences effect sizes from objective cognitive measures, and in fact Bayesian statistics demonstrate strong evidence for the null hypothesis at the meta-analytic level. Whether this indicates that the current active control conditions being used cannot capture these effects or that placebo and other non-specific intervention effects are minimal in this literature remains an open question. In either case, our empirical findings challenge the assumption in the field that only studies with active controls should be interpreted (Melby-Lervåg et al. 2016; Shipstead et al. 2012; Simons et al. 2016). Although this view is well-intentioned and ostensibly reasonable, our data demonstrate that it might be premature to only consider studies with active controls as valid, at least until such time as the research community is able to develop a better understanding of the specific and non-specific mechanisms of cognitive training in order to strategically design active controls that can control for the non-specific effects.

Finally, we contend that our results are straightforward, transparent, and replicable. Meta-analyses are often fraught with many complex, subjective decisions to be made, leading different researchers to arrive at different conclusions even when evaluating the same pool of studies (e.g., Au et al. 2015, 2016; Melby-Lervåg and Hulme 2016). Sympathetic to this issue, we endeavored to reduce the number of subjective decisions we made in order to get a relatively unbiased estimate of the active/passive control difference in cognitive training studies. To this end, we restricted our analyses to variables that were well defined and allowed reasonably straightforward coding. In instances of ambiguity, such as the issue of some researchers using active controls that closely resembled or were identical to experimental training tasks used by other researchers, we always relied on the authors' interpretations. Furthermore, we also point out that our double-controlled meta-analysis is in the rather unique position of being theoretically free of systematic publication bias since none of the included primary studies were published based on the merits of their control groups, and indeed, we found no evidence that publication bias differentially affects comparisons with active controls as comparisons with passive controls.

Funding This work was supported in part by the National Science Foundation Graduate Research Fellowship, Grant No. DGE-1321846 (JA) and the National Institute of Health/National Institute on Aging, Grant Nos. 1R01MH111742 and 1K02AG054665 (SMJ).

Compliance with Ethical Standards

Conflict of Interest MB and KB are employed at the MIND Research Institute, whose interests are related to this work. SMJ has an indirect financial interest in the MIND Research Institute. No other authors declare any competing financial interests.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: the same or different constructs? *Psychological Bulletin, 131*(1), 30–60. https://doi.org/10.1037/0033-2909.131.1.30.

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: a meta-analysis. *Psychonomic Bulletin & Review, 22*(2), 366–377. https://doi.org/10.3758/s13423-014-0699-x.

Au, J., Buschkuehl, M., Duncan, G. J., & Jaeggi, S. M. (2016). There is no convincing evidence that working memory training is NOT effective: a reply to Melby-Lervåg and Hulme (2015). *Psychonomic Bulletin & Review, 23*(1), 331–337. https://doi.org/10.3758/s13423-015-0967-4.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612–637.

Beecher, H. K. (1955). The powerful placebo. *Journal of the American Medical Association, 159*(17), 1602–1606.

Benedetti, F., Mayberg, H. S., Wager, T. D., Stohler, C. S., & Zubieta, J.-K. (2005). Neurobiological mechanisms of the placebo effect.

Journal of Neuroscience, 25(45), 10390–10402. https://doi.org/10.1523/JNEUROSCI.3458-05.2005.

Boot, W. R., Kramer, A. F., Simons, D. J., Fabiani, M., & Gratton, G. (2008). The effects of video game playing on attention, memory, and executive control. *Acta Psychologica, 129*(3), 387–398. https://doi.org/10.1016/j.actpsy.2008.09.005.

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*(4), 445–454. https://doi.org/10.1177/1745691613491271.

Borenstein, M. (2009). *Introduction to meta-analysis*. Chichester: Wiley.

Borenstein, M., Higgins, J., & Rothstein, H. (2005). *Comprehensive meta-analysis version 2 (version 2)*. Englewood: Biostat.

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods, 8*(1), 5–18. https://doi.org/10.1002/jrsm.1230.

Buschkuehl, M., Hernandez-Garcia, L., Jaeggi, S. M., Bernard, J. A., & Jonides, J. (2014). Neural effects of short-term training on working memory. *Cognitive, Affective, & Behavioral Neuroscience, 14*(1), 147–160. https://doi.org/10.3758/s13415-013-0244-9.

Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: a 40-year meta-analysis. *Psychological Bulletin, 140*(4), 980–1008. https://doi.org/10.1037/a0035661.

Chein, J. M., & Morrison, A. B. (2010). Expanding the mind's workspace: training and transfer effects with a complex working memory span task. *Psychonomic Bulletin & Review, 17*(2), 193–199. https://doi.org/10.3758/PBR.17.2.193.

Cleophas, T. J., & Zwinderman, A. H. (2017). Meta-meta-analysis. In T. J. Cleophas & A. H. Zwinderman (Eds.), *Modern meta-analysis: review and update of methodologies* (pp. 135–143). https://doi.org/10.1007/978-3-319-55895-0_11.

Dahlin, E., Neely, A. S., Larsson, A., Bäckman, L., & Nyberg, L. (2008). Transfer of learning after updating training mediated by the striatum. *Science, 320*(5882), 1510–1512. https://doi.org/10.1126/science.1155466.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629.

Foroughi, C. K., Monfort, S. S., Paczynski, M., McKnight, P. E., & Greenwood, P. M. (2016). Placebo effects in cognitive training. *Proceedings of the National Academy of Sciences, 113*(27), 7470–7474. https://doi.org/10.1073/pnas.1601243113.

Green, M. W., Taylor, M. A., Elliman, N. A., & Rhodes, O. (2001). Placebo expectancy effects in the relationship between glucose and cognition. *British Journal of Nutrition, 86*(02), 173. https://doi.org/10.1079/BJN2001398.

Green, C. S., Bavelier, D., Kramer, A. F., Vinogradov, S., Ansorge, U., Ball, K. K., et al. (2019). Improving methodological standards in behavioral interventions for cognitive enhancement. *Journal of Cognitive Enhancement.* https://doi.org/10.1007/s41465-018-0115-y.

Hróbjartsson, A., & Gøtzsche, P. C. (2001). Is the placebo powerless? *The New England Journal of Medicine, 2001*(344), 1594–1602.

Hróbjartsson, A., & Gøtzsche, P. C. (2004). Is the placebo powerless? Update of a systematic review with 52 new randomized trials comparing placebo with no treatment. *Journal of Internal Medicine, 256*(2), 91–100.

Hróbjartsson, A., & Gøtzsche, P. C. (2010). Placebo interventions for all clinical conditions. The Cochrane Library. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD003974.pub3/full. Accessed 1 June 2017.

Hsu, N. S., Buschkuehl, M., Jonides, J., & Jaeggi, S. M. (2013). *Potential mechanisms underlying working memory training and transfer. Presented at the Psychonomic Society Annual Meeting*. Toronto: Ontario.

Hsu, N. S., Jaeggi, S. M., & Novick, J. M. (2017). A common neural hub resolves syntactic and non-syntactic conflict through cooperation with task-specific networks. *Brain and Language, 166*, 63–77. https://doi.org/10.1016/j.bandl.2016.12.006.

Hussey, E. K., Harbison, J. I., Teubner-Rhodes, S. E., Mishler, A., Velnoskey, K., & Novick, J. M. (2017). Memory and language improvements following cognitive control training. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 43*(1), 23–58. https://doi.org/10.1037/xlm0000283.

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, 105*(19), 6829–6833. https://doi.org/10.1073/pnas.0801268105.

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

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving, 7*(1). https://doi.org/10.7771/1932-6246.1167.

Katz, B., Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (2018). The effect of monetary compensation on cognitive training outcomes. *Learning and Motivation, 63*, 77–90. https://doi.org/10.1016/j.lmot.2017.12.002.

Kirsch, I. (2005). Placebo psychotherapy: synonym or oxymoron? *Journal of Clinical Psychology, 61*(7), 791–803. https://doi.org/10.1002/jclp.20126.

Klauer, K. J., & Phye, G. D. (2008). Inductive reasoning: a training approach. *Review of Educational Research, 78*(1), 85–123. https://doi.org/10.3102/0034654307313402.

Lee, C. H., Cook, S., Lee, J. S., & Han, B. (2016). Comparison of two meta-analysis methods: inverse-variance-weighted average and weighted sum of Z-scores. *Genomics & Informatics, 14*(4), 173–180. https://doi.org/10.5808/GI.2016.14.4.173.

Li, D., Christ, S. E., & Cowan, N. (2014). Domain-general and domain-specific functional networks in working memory. *NeuroImage, 102*(02), 646–656. https://doi.org/10.1016/j.neuroimage.2014.08.028.

Looby, A., & Earleywine, M. (2011). Expectation to receive methylphenidate enhances subjective arousal but not cognitive performance. *Experimental and Clinical Psychopharmacology, 19*(6), 433–444. https://doi.org/10.1037/a0025252.

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental Psychology, 49*(2), 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*(1), 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*(4), 512–534.

Moher, D. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA statement. *Annals of Internal Medicine, 151*(4), 264. https://doi.org/10.7326/0003-4819-151-4-200908180-00135.

Morey, D., Rouder, J. N., & Jamil, T. (2014). Bayes factor: computation of Bayes factors for common designs (R package version 0.9.8). Retrieved from http://CRAN.R-project.org/package=BayesFactor. Accessed 1 Aug 2017.

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11*(2), 364–386. https://doi.org/10.1177/1094428106291059.

Nichols, A. L., & Maner, J. K. (2008). The good-subject effect: investigating participant demand characteristics. The Journal of general psychology, 135(2), 151-166.

Oelhafen, S., Nikolaidis, A., Padovani, T., Blaser, D., Koenig, T., & Perrig, W. J. (2013). Increased parietal activity after training of interference control. *Neuropsychologia, 51*(13), 2781–2790. https://doi.org/10.1016/j.neuropsychologia.2013.08.012.

Opitz, B., Schneiders, J. A., Krick, C. M., & Mecklinger, A. (2014). Selective transfer of visual working memory training on Chinese character learning. *Neuropsychologia, 53*, 1–11. https://doi.org/10.1016/j.neuropsychologia.2013.10.017.

Pahor, A., Jaeggi, S. M., & Seitz, A. R. (2018). Brain training. In *ELS* (pp. 1–9). https://doi.org/10.1002/9780470015902.a0028037.

Price, D. D., Finniss, D. G., & Benedetti, F. (2008). A comprehensive review of the placebo effect: recent advances and current thought. *Annual Review of Psychology, 59*(1), 565–590. https://doi.org/10.1146/annurev.psych.59.113006.095941.

R Core Team. (2013). R: a language and environment for statistical computing. Retrieved from http://www.R-project.org/

Rass, O., Forsyth, J. K., Bolbecker, A. R., Hetrick, W. P., Breier, A., Lysaker, P. H., & O'Donnell, B. F. (2012). Computer-assisted cognitive remediation for schizophrenia: a randomized single-blind pilot study. *Schizophrenia Research, 139*(1–3), 92–98. https://doi.org/10.1016/j.schres.2012.05.016.

Rebok, G. (2015). Selecting control groups: to what should we compare behavioral interventions? In L. N. Gitlin & S. J. Czaja (Eds.), *Behavioral intervention research: designing, evaluating, and implementing* (pp. 139–160). New York, NY: Springer Publishing Company.

Riley, R. D., Higgins, J. P. T., & Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *BMJ, 342*. https://doi.org/10.1136/bmj.d549.

Rohatgi, A. (2017). WebPlotDigitizer Version 3.12 (Version 3.12). Retrieved from http://arohatgi.info/WebPlotDigitizer. Accessed 1 Dec 2016.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (1st ed.). Newbury Park: SAGE Publications, Inc..

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225.

Salmi, J., Nyberg, L., & Laine, M. (2018). Working memory training mostly engages general-purpose large-scale networks for learning. *Neuroscience and Biobehavioral Reviews, 93*, 108–122. https://doi.org/10.1016/j.neubiorev.2018.03.019.

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*. https://doi.org/10.3389/fnagi.2010.00027.

Schneiders, J. A., Opitz, B., Krick, C. M., & Mecklinger, A. (2011). Separating intra-modal and across-modal training effects in visual working memory: an fMRI investigation. *Cerebral Cortex, 21*(11), 2555–2564. https://doi.org/10.1093/cercor/bhr037.

Schwarz, K., & Büchel, C. (2015). Cognition and the placebo effect—dissociating subjective perception and actual performance. *PLoS One, 10*(7), 1–12. https://doi.org/10.1371/journal.pone.0130492.

Schweizer, K. (2007). Investigating the relationship of working memory tasks and fluid intelligence tests by means of the fixed-links model in considering the impurity problem. *Intelligence, 35*(6), 591–604. https://doi.org/10.1016/j.intell.2006.11.004.

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective? *Psychological Bulletin, 138*(4), 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*(3), 103–186. https://doi.org/10.1177/1529100616661983.

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

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*, 1–20. https://doi.org/10.3758/s13423-016-1217-0.

Stephenson, C. L., & Halpern, D. F. (2013). Improved matrix reasoning is limited to training on tasks with a visuospatial component. *Intelligence, 41*(5), 341–357. https://doi.org/10.1016/j.intell.2013.05.006.

Thompson, T. W., Waskom, M. L., Garel, K.-L. A., Cardenas-Iniguez, C., Reynolds, G. O., Winter, R., et al. (2013). Failure of working memory training to enhance cognition or intelligence. *PLoS One, 8*(5), e63614. https://doi.org/10.1371/journal.pone.0063614.

Thorell, L. B., Lindqvist, S., Bergman Nutley, S., Bohlin, G., & Klingberg, T. (2009). Training and transfer effects of executive functions in preschool children. *Developmental Science, 12*(1), 106–113. https://doi.org/10.1111/j.1467-7687.2008.00745.x.

Tsai, N., Buschkuehl, M., Kamarsu, S., Shah, P., Jonides, J., & Jaeggi, S. M. (2018). (Un)great expectations: the role of placebo effects in cognitive training. *Journal of Applied Research in Memory and Cognition*. https://doi.org/10.1016/j.jarmac.2018.06.001.

Vartanian, O., Coady, L., & Blackler, K. (2016). 3D multiple object tracking boosts working memory span: implications for cognitive training in military populations. *Military Psychology, 28*(5), 353–360. https://doi.org/10.1037/mil0000125.

Wager, T. D., & Atlas, L. Y. (2015). The neuroscience of placebo effects: connecting context, learning and health. *Nature Reviews. Neuroscience, 16*(7), 403–418. https://doi.org/10.1038/nrn3976.

Weicker, J., Villringer, A., & Thöne-Otto, A. (2016). Can impaired working memory functioning be improved by training? A meta-analysis with a special focus on brain injured patients. *Neuropsychology, 30*(2), 190–212. https://doi.org/10.1037/neu0000227.

Willis, S. L. (2001). Methodological issues in behavioral intervention research with the elderly. In *Handbook of the psychology of aging* (5th ed., pp. 78–108). San Diego: Academic Press.