Psychon Bull Rev
DOI 10.3758/s13423-014-0699-x

THEORETICAL REVIEW

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

Jacky Au · Ellen Sheehan · Nancy Tsai · Greg J. Duncan · Martin Buschkuehl · Susanne M. Jaeggi

Abstract Working memory (WM), the ability to store and manipulate information for short periods of time, is an important predictor of scholastic aptitude and a critical bottleneck underlying higher-order cognitive processes, including controlled attention and reasoning. Recent interventions targeting WM have suggested plasticity of the WM system by demonstrating improvements in both trained and untrained WM tasks. However, evidence on transfer of improved WM into more general cognitive domains such as fluid intelligence (*Gf*) has been more equivocal. Therefore, we conducted a meta-analysis focusing on one specific training program, *n*-back. We searched PubMed and Google Scholar for all *n*-back training studies with *Gf* outcome measures, a control group, and healthy participants between 18 and 50 years of age. In total, we included 20 studies in our analyses that met our criteria and found a small but significant positive effect of *n*-back training on improving *Gf*. Several factors that moderate this transfer are identified and discussed. We conclude that short-term cognitive training on the order of weeks can result in beneficial effects in important cognitive functions as measured by laboratory tests.

Keywords Cognitive training · Transfer · Plasticity

Electronic supplementary material The online version of this article (doi:10.3758/s13423-014-0699-x) contains supplementary material, which is available to authorized users.

J. Au (✉) · E. Sheehan · N. Tsai · G. J. Duncan · S. M. Jaeggi
School of Education, University of California, Irvine, Irvine, CA 92697, USA
e-mail: jwau@uci.edu

M. Buschkuehl
MIND Research Institute, Irvine, CA, USA

Published online: 08 August 2014

Introduction

"It is becoming very clear that training on working memory with the goal of trying to increase *Gf* will likely not succeed." (Harrison, Shipstead, Hicks, Hambrick, Redick, & Engle, 2013, p. 2418)

The cognitive training literature has seen an explosion of recent interest in exploring the claim that gains in working memory (WM) training might transfer to gains in measures of fluid intelligence (*Gf*). If true, the implications for academic, professional, and personal success are considerable (Gottfredson, 1997). Despite many promising studies (e.g., Jaeggi, Buschkuehl, Jonides, & Perrig, 2008; Rudebeck, Bor, Ormond, O'Reilly, & Lee, 2012; Stephenson & Halpern, 2013), the aforementioned quote reflects results from other well-controlled, rigorous attempts at replication that have failed to find transfer (Redick et al., 2013; Thompson et al., 2013). Thus, the debate continues without consensus. However, disparate replication results may be the product of disparate conditions, some of which facilitate transfer and others of which impede it. Without careful consideration of these conditions, any categorical claim about positive or negative impacts is premature (Jaeggi, Buschkuehl, Shah, & Jonides, 2014). Therefore, we conducted a systematic meta-analysis of the entire extant literature in order to estimate an overall average effect size and to explore moderators associated with deviations from the overall average.

The debate over the malleability of intelligence is deeply rooted in the history of psychology, stemming as far back as the late 19th century, when Francis Galton promoted his views on the strict heritability of intelligence (Galton, 1892). Despite popular critics such as Alfred Binet (1909), forefather of modern IQ testing, research over the next few decades brought on a zeitgeist of determinism that pervaded popular scientific thought, borne out by work on developmental critical periods

and twin studies of inheritance. This view of the immutability of intelligence came to influence more modern thinkers (Caruso, Taylor, & Detterman, 1982) and was reinforced by the difficulties many researchers faced in demonstrating transfer (Detterman, 1993; Salomon & Perkins, 1989). But more recent evidence has accumulated suggesting malleability of intelligence, including research demonstrating the positive impacts of such interventions as exercise, education, nutrition, and even the industrialization of nations (Dickens & Flynn, 2001; Nisbett et al., 2012). Most scientists today acknowledge the importance of genetics and heritability in the development of intelligence, while recognizing the role that certain environments can play in molding this development.

One of the critical components of general intelligence is *Gf,* or the ability to reason in novel situations independently of previous knowledge. It is the aspect of intelligence that has shown the greatest malleability over time in the documented Flynn effect (Dickens & Flynn, 2001), as well as in experimental work (e.g., Stankov, 1986). It is also highly predictive of professional and educational success (Gottfredson, 1997) and has, therefore, been a prime target of intervention. One of the core processes driving *Gf*, as well as other higher cognitive abilities, is WM (Wiley, Jarosz, Cushen, & Colflesh, 2011). Estimates report a shared variance of at least 50% (Kane, Hambrick, & Conway, 2005; Oberauer, Schulze, Wilhelm, & Suss, 2005), and neuroimaging evidence has demonstrated functional overlap in the lateral prefrontal and parietal cortices, implicating shared neural resources that underlie both constructs (Burgess, Gray, Conway, & Braver, 2011). This makes sense, since any attempt to reason through a novel situation requires maintaining multiple possible goals in WM while simultaneously manipulating that information in order to achieve the desired goal.

The recent prospect of improving WM through training has raised the possibility of concomitant improvements in *Gf* (cf. von Bastian & Oberauer, 2013). More specifically, the *n*-back task, which requires not only the storage and continual updating of information in WM, but also interference resolution, has been used widely in WM training studies that explore transfer to *Gf.* The *n*-back task involves serial presentation of a stimulus (e.g., a shape), spaced several seconds apart. The participant must decide whether the current stimulus matches the one displayed *n* trials ago, where *n* is a variable number that can be adjusted up or down to respectively increase or decrease cognitive load. In the context of WM training, efforts have focused on flexibly adapting the task difficulty in accordance with the participant's fluctuating performance level by increasing and decreasing the level of *n*. The idea is to keep the participant's WM system perpetually engaged at its limit, thereby stimulating an increase in WM function, which may then translate into more general improvements in tasks that rely on the integrity of WM skills, such as *Gf* (Jaeggi et al., 2008).

Despite some initial successes (cf. Buschkuehl & Jaeggi, 2010), it still remains contentious that intellectual plasticity can continue beyond early developmental periods and into adulthood, particularly as a result of relatively brief cognitive training on the order of weeks with a simple, repetitive computerized intervention such as *n*-back. However, we are now at a point in this debate where the publication of a critical mass of WM training studies based on just one type of intervention (*n*-back) and just one type of population (healthy, young adults) warrants a meta-analysis that can inform this debate in ways that simple vote-counting procedures of systematic reviews cannot.

The purpose of the present meta-analysis is twofold: first, to estimate the net effect size of *Gf* improvement in healthy, young adults as a function of *n*-back training and, second, to elucidate factors that may moderate this transfer. For example, building off previous work examining individual differences underlying training outcomes, we hypothesized that remuneration for study participation could dampen the training effect by reducing intrinsic motivation (Jaeggi et al., 2014). This idea is based on a broader literature demonstrating negative effects of extrinsic rewards (e.g., money) on intrinsic motivation (Deci, Koestner, & Ryan, 1999). Moreover, remuneration could also bias the recruitment process by preferentially attracting participants more invested in money than in self-improvement, a quality that may preclude the type of "grit" (Duckworth, Peterson, Matthews, & Kelly, 2007) needed for successful training (Jaeggi et al., 2014). We therefore hypothesized an inverse correlation between amount of payment and *Gf* improvement.

We also sought to empirically test several claims that have been made in the literature. For example, many studies use passive (or no-contact) control groups that receive no intervention. While this does control for important test–retest practice effects, it does not control for potential motivational or Hawthorne effects associated with being enrolled in an intervention study (Shipstead, Redick, & Engle, 2012). Therefore, the training-based improvements seen in studies using passive controls might be due to nonspecific effects such as expectations of improvement or heightened motivation.

With this in mind, two recent, widely publicized training studies used active controls and failed to find transfer to measures of *Gf* (Harrison et al., 2013; Redick et al., 2013). It is important to note, however, that their active control groups did not improve over baseline, nor did they outperform their associated passive control groups, suggesting that the failure to find transfer was irrespective of control type. Therefore, we tested the effect of control type across studies. Furthermore, we have previously described a dose-dependent relationship between training and transfer, such that more training leads to more gain (Jaeggi et al., 2008). This relationship has been replicated in some studies (Basak, Boot, Voss, & Kramer,

2008; Dahlin, Backman, Neely, & Nyberg, 2009; Stepankova et al., 2014; Tomic & Klauer, 1996), but not others (Redick et al., 2013). We hypothesized that this dosage effect would hold up across studies in the form of a positive correlation between number of training sessions and degree of transfer.

Finally, due to the broad interest in cognitive training, laboratories around the world are investigating the effects of training and transfer. In fact, the first study of *n*-back training on *Gf* was conducted in Switzerland (Jaeggi et al., 2008), and from our own experiences conducting research both internationally and in the U.S., we have anecdotally observed motivational differences across cultures. Therefore, we sought to systematically test for any regional differences in *Gf* gains. Such differences would not be surprising given the growing literature demonstrating culturally mediated effects on various aspects of cognitive functioning, including attention and learning (Ketay, Aron, & Hedden, 2009; Muggleton & Banissy, 2014; Nisbett & Norenzayan, 2002). Along a similar vein, we further tested for research laboratory effects by estimating whether transfer rates among studies involving the original *n*-back training researchers, S.M.J. and M.B., differed from average transfer rates from other laboratories. Other potential moderators of interest are summarized further below (c.f., Methods).

Method

Study selection

We searched the PubMed and Google Scholar databases using the following keywords taken separately or in combination: *n*-back training, WM training, cognitive training, fluid intelligence. Several unpublished dissertations were also found on Google Scholar by incorporating the keyword "dissertation" or "thesis" into one of the above search terms. We also included unpublished work from researchers known to us. Finally, we checked the references of selected papers and searched relevant conference proceedings that were accessible to us in order to ensure that there were not any additional studies we had omitted.

Our inclusion criteria were as follows: Studies must have trained participants on some form of adaptive *n*-back, included a control group, and used some form of *Gf* outcome measure. In order to avoid the confounding effects of development and senescence, we restricted our analysis to healthy young adults between the ages of 18 and 50 years. Studies using a battery of different training interventions where the effects of *n*-back could not be isolated were excluded. Similarly, studies with missing or incomplete data relevant to our effect size calculations were also excluded, as were studies that trained for too short a duration (less than 1 week). In the end, 20 studies remained in the final meta-analysis (Fig. 2). All papers were written in the English language. Study selection criteria are detailed in Fig. 1.

Coding

After study selection was completed, coding commenced independently by two small teams. S.M.J. and M.B. made up one team, while J.A. and E.S. made up another. Percent agreement on the coding (interrater reliability) was high (94.14%), with $\kappa = 0.88$ ($SE = 0.03$) using a conservative expected agreement estimate of 50% (i.e., match or no match). This falls in the "almost perfect agreement" range (Viera & Garrett, 2005). Any disagreements were discussed and resolved as a group. Thirty distinct treatment groups and 24 distinct control groups were identified, leading to 24 group comparisons. Where multiple control groups existed within a study, such as an active and passive control, the active control was chosen, provided that the control intervention did not load on WM or some other process that might itself improve *Gf*. For example, Stephenson and Halpern (2013) used a spatial span active control task that also tapped WM, and Oelhafen et al. (2013) investigated the use of lure trials in *n*-back and, therefore, considered an adaptive *n*-back without lures to be an active control. In both cases, the passive control results were selected.

The primary outcomes of interest were treatment/control differences in scores on the *Gf* tasks used by each study. *Gf* is typically defined as the ability to think logically and reason through problems in novel situations, independently from previously acquired knowledge. In our selection of *Gf* tasks, we followed previously published guidelines that include a list of common metrics that load strongly on *Gf* (Ackerman, Beier, & Boyle, 2005; Gray & Thompson, 2004). All tasks selected in our meta-analysis are taken directly from this list or are similar in construct. In the end, all authors reviewed and agreed upon the *Gf* classifications in this article, and they are all included in the supplementary online materials (SOM; Table S3).

Another variable of interest was remuneration for participating in a training study. This was reported directly in the papers as either a lump sum or an hourly rate. In the latter case, an estimate had to be made on the basis of the duration of study participation. When remuneration was not reported, the authors were contacted directly for the information. For international studies, remuneration was coded in U.S. dollars based on the exchange rate at the time of submission for that particular article. All values were inflated from the date of publication to current U.S. price levels in 2014.

Furthermore, we quantified several different dimensions of the training regimen, such as number of sessions (days) and length per session (minutes). We also modeled the training curves published in individual studies using a regression

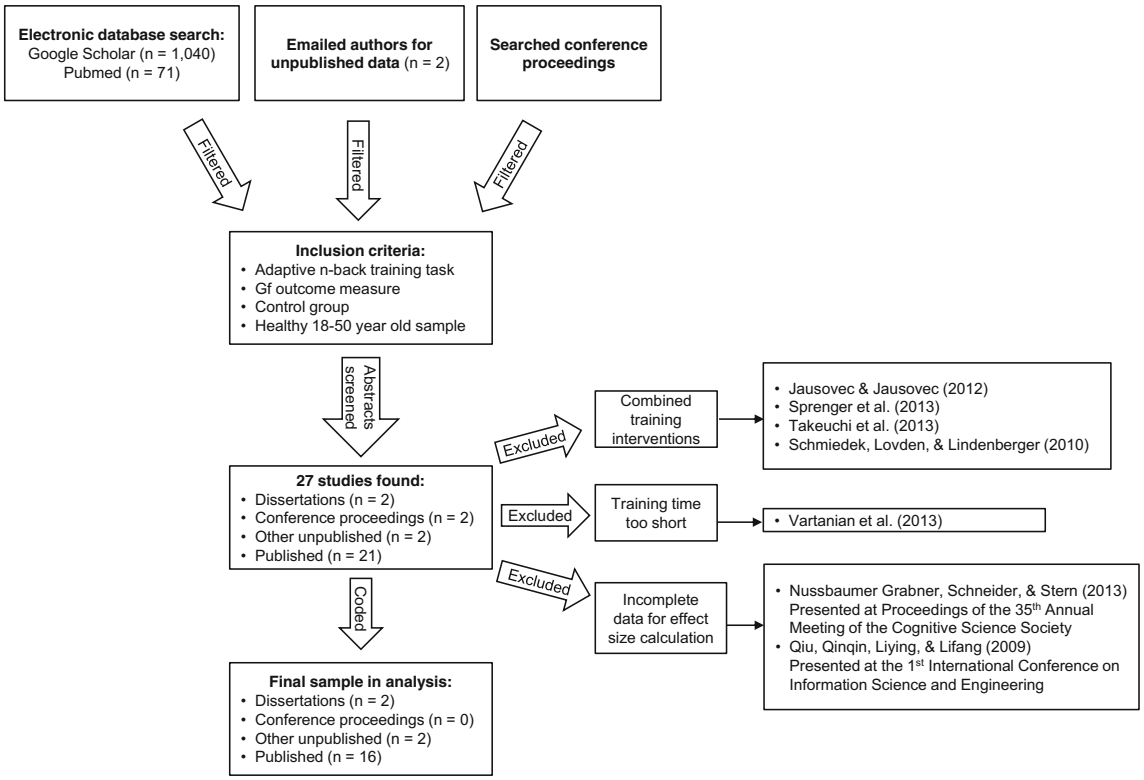

Fig. 1 Overview of study selection criteria

approach with logarithmic transformation of session number. This analysis yielded two parameters: the starting *n*-back level on day 1 (intercept) and the rate of improvement (slope). To ensure comparability, only dual *n*-back studies were included in this training curve analysis ($N = 20$), although two studies were excluded because their session lengths were twice as long as those in most other studies and, thus, their training curves were deemed incomparable (Kundu, Sutterer, Emrich, & Postle, 2013; Thompson et al., 2013). There were not enough single *n*-back studies to perform a separate analysis: Of the seven single *n*-back training groups, four had incomplete information (Schwarb, 2012; Stephenson & Halpern, 2013), and one (Heinzel et al., 2014) was excluded due to a different adaptivity algorithm, leaving only two remaining studies, which we did not analyze.

Other variables of interest that we coded were type of control group (active or passive), training location (international or U.S., laboratory or home), *n*-back type (dual or single, visual or auditory), whether or not our team was involved, type of *Gf* measure (matrix reasoning or other, visual or verbal), attrition rate, and age.

Statistical analyses

We first assessed publication bias, also referred to as the "file drawer problem," in our sample of studies (Sterne, Gavaghan, & Egger, 2000). This refers to the phenomenon in which studies reporting null results are less likely to be published and, therefore, the extant literature included in a meta-analysis is susceptible to bias. Although we attempted to address this issue by including unpublished works within our data set, it is likely that there are more out there. Publication bias can never be fully resolved, only mitigated. However, in our case, it is fortunate that the controversy surrounding WM training has facilitated the publication of null effects, which may result in a more representative sample of impact estimates. Nevertheless, statistical methods were used to estimate the extent of bias in our sample.

Next, we assessed heterogeneity using the I^2 statistic, which represents the percentage of total variation between studies that is due to heterogeneity rather than chance or sampling error alone (Higgins, Thompson, Deeks, & Altman, 2003). In other words, a high I^2 value implies that not all of the studies included in the meta-analysis are, in fact, measuring the same effect and that some of the variation is due to differences in study design or sampling biases across studies, rather than sampling error alone. Conversely, a low I^2 value indicates homogeneity across studies and argues for the generalizability of results.

Effect sizes (ESs) were calculated with the Comprehensive Meta-Analysis (CMA) software package (Borenstein, Higgins, & Rothstein, 2005), using a random effects model to calculate standardized mean differences (SMDs), adjusted for small sample sizes using Hedges *g* (Rosenthal, 1991). This was done in two different ways. First, we calculated the SMD

between posttest and pretest for both the treatment (Tx ES) and control (Ctrl ES) groups separately. Positive values indicate improvements on *Gf* measures at posttest, relative to pretest. Additionally, we also calculated the SMDs between treatment and control groups at posttest only (Post ES), in order to summarize each study with only one ES (and its associated standard error) that directly compares treatment and control groups. Positive values indicate improvements in treatment groups above and beyond improvements in control groups. Post ESs were calculated per previous recommendations (Dunst, Hamby, & Trivette, 2004; Higgins, & Green, 2011), under the assumption that no group differences existed at pretest. This assumption was tested by looking at the SMD between treatment and control groups at pretest (baseline ES). For simplicity, subsequent interpretations, unless otherwise specified, are based on Post ES, but Tx ES and Ctrl ES calculations are also presented in Table 1 for comparison. Variances of ES were tested for equality by Levene's test, and *t*-tests were conducted accordingly to determine significance levels of ES comparisons.

Moderator analyses were performed with both categorical and continuous predictors. We divided our studies into subgroups based on categorical predictors (e.g., dual vs. single *n*-back), and compared individual subgroup ESs. Post ESs were regressed on continuous predictors, covarying out baseline differences, using CMA's built-in unrestricted maximum likelihood meta-regression function. Some of our data were influenced by outlying values, but it was not always clear whether the retention or removal of outliers made more theoretical sense. Therefore, given the relatively small sample sizes and the large, but not necessarily undue influence that these outliers can exert, we have included alternate regression models in the SOM (Table S1) that have outliers trimmed for comparison. Outliers were categorized as data points that were 3 median absolute deviations away from the overall median (Leys, Ley, Klein, Bernard, & Licata, 2013).

ESs from multiple *Gf* outcomes within a single treatment or control group were averaged together into one net effect. Similarly, ESs from multiple treatment groups (e.g., dual and single *n*-back) within a single study were collapsed into one weighted average (based on sample size) if they were compared with the same control group. If each treatment group had its own control group within a study, then ESs were calculated separately and treated as independent (Borenstein, Hedges, Higgins, & Rothstein, 2009). The end result was an independent set of ESs such that each treatment or control group was never represented more than once in the overall analysis ($n = 24$).

In subgroup analyses, however, a single treatment or control group may be factored into the separate ES calculations of different subgroups. For example, in Jaeggi et al. (2014), two treatment groups (dual and single *n*-back) were compared with the same control group. In the subgroup analysis of dual versus single *n*-back, this same control group was compared separately with each treatment group within the dual and single *n*-back subgroups (creating two separate ESs), even though the overall analysis collapsed these three groups into just one net ES. This explains some of the sample size discrepancies in Table 1 (i.e., the sum of subgroup sample sizes can be greater than the overall sample size of 24). In the end, the same data were never represented more than once in any one particular ES calculation, creating an independent set of ESs in each of our analyses.

Results

Description of studies

The 20 studies included here were all completed between 2008 and 2013. Sample sizes of treatment groups varied between 7 and 36 participants (mean ± *SD*: 19.96 ± 8.13), and control groups between 8 and 43 (mean ± *SD*: 19.29 ± 8.74). Mean age of participants was 22.85 years (*SD*: 2.60). In total, we analyzed data from 98 *Gf* outcome measures among 559 *n*-back trained participants and 463 controls. See SOM (Table S3) for a list of each of the unique *Gf* outcomes used in our meta-analysis.

A statistical analysis revealed no evidence of publication bias in our sample of studies. Egger's regression (Egger, Davey Smith, Schneider, & Minder, 1997), which is based on the association between standard error and ES, revealed no relationship ($p = .68$). Therefore, smaller studies (indexed by higher standard errors) are not systematically reporting higher ESs (Fig. 2), as would be expected in the presence of publication bias, since smaller studies are more likely to show extreme ESs. Nevertheless, since our relatively small sample of studies lack substantial power, we also calculated the classic fail-safe *N* test (Orwin, 1983), which revealed that it would take 59 studies reporting null results ($g = 0$) to be included in our analyses in order for our findings to lose statistical significance.[1]

We also estimated heterogeneity, which quantifies the between-study variation caused by differences other than sampling error (e.g., study design, etc.). Some heterogeneity ($I^2 = 27.88\%$, $p = .08$) was found in Tx ES, which only looked at improvements within the treatment groups, but this heterogeneity was largely controlled out in Post ES ($I^2 = 6.92\%$, $p = .37$), which directly compares treatment to control groups at posttest. According to the guidelines laid out by Higgins et al. (2003), values of 25%, 50%, and 75% are considered low, moderate, and high amounts of heterogeneity, respectively.

[1] Classic fail-safe *N* calculation: The 24 observed studies, with weighted $g = 0.24$, led to a combined weighted ES of $g = 5.78$. Adding the 59 null effect studies would increase the sample size to $n = 83$. Average ES in the new sample would be: $\frac{5.78}{83} = .069$. The new SE would be calculated assuming the same *SD*: $\frac{0.338}{(\sqrt{83})} = .037$. Therefore, with 59 null effect studies, the ES would drop down to $g = .07$, with $SE = .04$.

Table 1 Overall and subgroup analyses

Overall	***n***	**ES**	***p***						
Post ES	24	.24 (.07)	.03*						
Baseline ES	24	−.003 (.08)							
Tx ES	30	.41 (.07)	.03*						
Ctrl ES	24	.18 (.07)							
Subgroups	***n***	**Post ES**	***p***	***n***	**Tx ES**	***p***	***n***	**Ctrl ES**	***p***
Active Control	12	.06 (.09)	.01*	13	.25 (.10)	.04*	12	.08 (.10)	.20
Passive Control	12	.44 (.10)		17	.54 (.90)		12	.28 (.10)	
International	13	.44 (.10)	.01*	15	.66 (.09)	<.01*	13	.29 (.09)	.13
U.S.	11	.06 (.09)		15	.21 (.08)		11	.08 (.09)	
Laboratory	16	.23 (.09)	.81	20	.36 (.09)	.32	16	.22 (.08)	.50
Home	8	.27 (.12)		10	.51 (.12)		8	.11 (.12)	
Dual *n*-Back	22	.24 (.07)	.53	23	.43 (.08)	.69	22	.18 (.07)	.86
Single *n*-Back	5	.36 (.13)		7	.36 (.14)		5	.15 (.13)	
Auditory *n*-Back	3	.27 (.16)	.46	3	.21 (.20)	.29	3	.01 (.16)	.64
Visual *n*-Back	6	.46 (.13)		6	.55 (.16)		6	.14 (.12)	
Visual and auditory *n*-Back[¥]	20	.21 (.08)		21	.41 (.09)		20	.19 (.07)	
Jaeggi/ Buschkuehl group	10	.23 (.11)	.89	12	.46 (.11)	.58	10	.24 (.10)	.52
Other research group	14	.25 (.10)		18	.38 (.09)		14	.14 (.09)	
Matrix *Gf* tasks	23	.20 (.08)	.53	29	.32 (.07)	.67	23	.05 (.07)	.10
Nonmatrix *Gf* Tasks[§]	13	.13 (.08)		17	.37 (.1)		13	.30 (.12)	
Verbal *Gf* tasks	5	.13 (.12)	.54	6	.33 (.18)	.70	5	.39 (.23)	.25
Visuospatial *Gf* tasks[§]	24	.23 (.07)		30	.40 (.07)		24	.15 (.07)	

Note. Table of all effect size calculations. Numbers in parentheses represent standard errors. Post effect size (ES) was calculated as standardized mean difference (SMD) between treatment and control at posttest. Baseline ES was SMD between treatment and control at baseline. Tx ES was SMD between pre- and posttests of the treatment groups. Ctrl ES was SMD between pre- and posttests of the control groups. The Post ES estimates of the international/ U.S. and active/passive control subgroups draw upon many of the same studies and, therefore, yield similar and significant results. However, the identical nature of the pairs of ES estimates is coincidence and partially an artifact of rounding. See Discussion.

[¥] In these analyses, visual and auditory *n*-back is a subset of dual *n*-back, which also includes dual visual modalities.

[+] *F* test from ANOVA revealed no significant differences between any of the three means.

[§] ES in these analyses are smaller than overall average, due to disaggregation of *Gf* measures in these calculations.

Our meta-analysis of *n*-back training studies in healthy young adults therefore shows an overall low to trivial amount of statistical heterogeneity. This makes sense, given our restricted analysis of only one intervention type (*n*-back) among only one subpopulation: healthy, young adults. However, the assumption of homogeneity may be premature. The relatively small sample sizes in our studies lead to wide, overlapping confidence intervals between studies (Fig. 2), which contribute toward the statistical assumption of homogeneity. In fact, the confidence intervals overlap almost entirely with the range of effect that *n*-back could reasonably be assumed to have (see overall effect size below), thereby underpowering the ability to detect heterogeneity. However, with ESs ranging from −0.28 to 1.11 (Fig. 2) and clear methodological differences in how *n*-back is implemented across studies, it is important to assess these differences with moderator analyses, which we have done further below.

Overall effect size

Results from individual studies are shown in Fig. 3 and detailed in Table 1. The treatment/control group difference in *Gf* at posttest ($g = 0.24$, $SE = 0.07$) is significantly greater than the treatment/control group difference at baseline ($g = -0.003$, $SE = 0.08$; group difference: $p = .03$). With a baseline ES of essentially 0, we conclude that no preexisting differences are present at the group level between treatment and control groups. The relatively large standard error of this difference (0.08) is due to two individual studies (Salminen, Strobach, & Schubert, 2012; Schweizer, Hampshire, & Dalgleish, 2011) that did show statistically significant but opposite-signed baseline differences. When we calculate the ES of pre- to posttest improvement separately for both treatment and control groups and take their difference, we obtain an almost identical overall effect (Tx ES – Ctrl ES = 0.41 – 0.18 = 0.23).

Subgroup analyses

Table 1 also shows the subgroup analyses for categorical moderators. Two subgroup contrasts reached conventional levels of statistical significance, with coincidentally identical pairs of ES estimates (see the Discussion section). First,

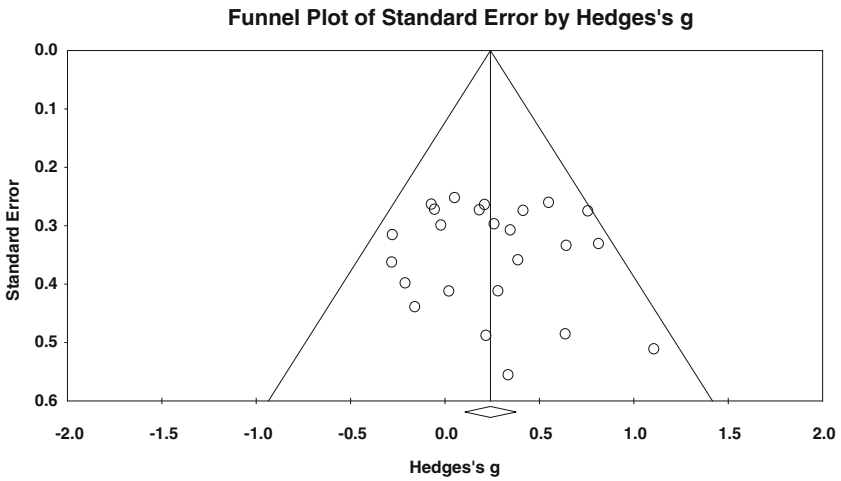

Fig. 2 Funnel plot of publication bias. Individual studies' mean weighted Post ES are graphed against standard error (index of sample size). Studies with the largest standard errors (smallest sample sizes) are shown at bottom

international (outside the U.S.) studies demonstrated an average ES of .44 (*SE* = .10), while U.S. studies only averaged .06 (*SE* = .09). The difference is significant at the $p < .01$ level. Second, studies that used passive controls demonstrated more net transfer ($g = 0.44$, *SE* = .10) than those with active controls ($g = 0.06$, *SE* = .09), a difference that is also statistically significant at $p < .01$. However, there was no difference in the performance of either type of control group when compared directly with each other (Ctrl ES; $p = .2$), but there was significant improvement in the performance of treatment groups (Tx ES; $p = .04$) within those studies that also use passive controls. Since Tx ES is calculated independently of the control group, the improvements found in these studies are irrespective of the type of control used.

Additionally, we found no difference depending on whether studies used dual or single *n*-back, which corroborates previous findings (Jaeggi et al., 2014; Jaeggi et al., 2010). We could not test the hypothesis that visual training is more effective than auditory training (cf. Stephenson & Halpern, 2013), owing to an inadequate sample size of 3 studies within

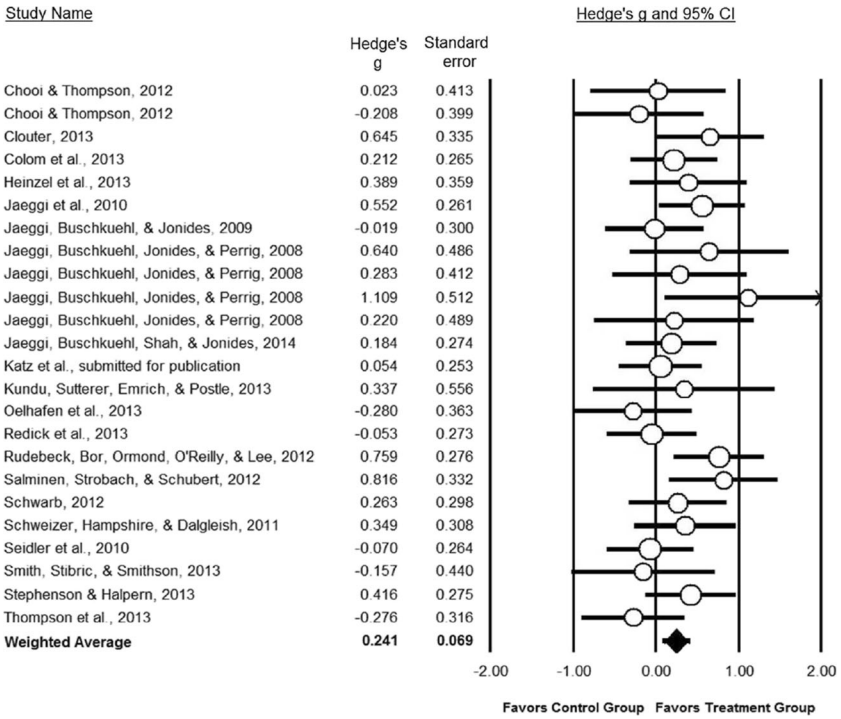

Fig. 3 Overall effect size across studies. Forest plot showing Post ES (Hedges *g*) and 95% confidence intervals from each individual study. Overall weighted ES is displayed at bottom

the auditory group. Training in the laboratory or at home did not make a difference either. We also assessed the nature of the transfer by examining whether certain types of *Gf* tests were more sensitive to *n*-back training than others. Transfer seemed to occur equally to both verbal and visual *Gf* tasks, although there were only a minority of studies that employed any form of verbal reasoning ($n = 5$), making this finding preliminary. Similarly, we saw fairly equal transfer both to matrix reasoning tasks (the most commonly used measure to assess *Gf* transfer) and to other types of *Gf* tasks, thereby suggesting that the transfer of *n*-back training may not be restricted to matrix reasoning tests (Shipstead et al., 2012) but, rather, may reflect a more global improvement in *Gf*. Finally, there was no statistical difference in ES between the 7 studies involving study authors S.J.M. and M.B. and the 13 studies that did not.

Regression analyses

Our regression models, as summarized in Table 2, revealed one primary finding: The effect of remuneration for study participation (in units of hundreds of dollars) was significantly and negatively associated with Post ES ($p = .05$, $b = -.07$), controlling for baseline differences (Fig. 4). However, the effect lost significance when outliers were trimmed ($p = .22$, $b = -.07$; SOM, Table S1), although the slope remains the same. No other predictors reached significance, but session length ($p = .06$, $b = -.03$) and starting *n*-back level ($p = .06$, $b = -.33$) both correlated marginally and negatively with Post ES when outliers were removed (SOM, Table S1). We also ran several multiple regression models (Table 3) to examine possible confounding variables that help explain the greater effect observed in studies with passive controls detailed above. Models 2 and 3 in Table 3 examined the differential effects of controlling for remuneration and international status (with baseline differences covaried out). Both individually appeared to contribute toward the effect observed in passive controls and, together, reduced the ES gain caused by the condition of having a passive control from .31 to .09, rendering the *p*-value nonsignificant.

Table 2 Regressions of continuous moderators

Moderators	*n*	*b* (*SE*)	*p*
Remuneration (hundreds of dollars)	24	−.07 (.03)	.05*
Session length (minutes)	24	−.01 (.01)	.21
No. of sessions	24	−.01 (.02)	.72
Starting *n*-back level	20	−.02 (.12)	.89
Rate of training improvement (slope)	20	−.45 (.42)	.29

Note. Regression table of Post ES on continuous moderators, covarying out baseline differences. Post ES is defined as standardized mean difference (Hedges *g*) between treatment and control groups at posttest. See Fig. 4 for graph of remuneration. Outliers, defined as 3 median absolute deviations from the overall median, are reported and trimmed in the SOM (Table S2).

* $p \leq .05$

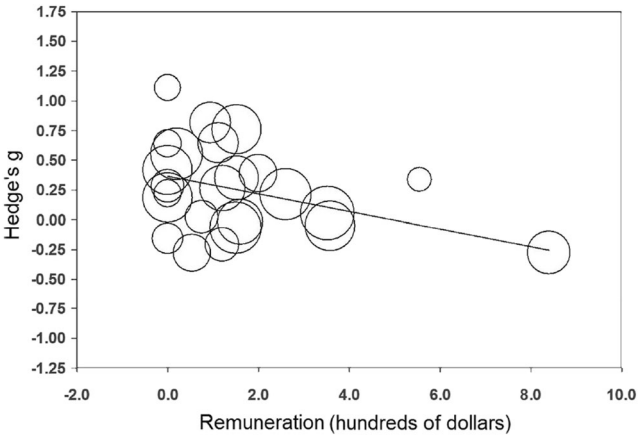

Fig. 4 Regression plot of Hedges *g* (Post ES) on remuneration. Diameters of circles are proportional to the sample size of the training groups they represent. Summary data are presented in Table 2

Discussion

The primary finding from this meta-analysis is a small but statistically significant net effect of *n*-back training on *Gf* outcome measures ($g = 0.24$, $SE = .07$), about the equivalent of 3–4 points on a standardized IQ test. Despite the small ES, several important points should be borne in mind. First is that these results are based on a restricted range of the general population: healthy, young adults between 18 and 50 years of age who were largely at the peak of cognitive functioning. Within this restricted range of young adults, the majority of participants were college undergraduates, thereby skewing our samples even younger (mean age ± *SD*: 22.85 ± 2.60). A common property of statistics holds that sampling from restricted ranges of the total population usually biases ES downward, due to reduced variability (Bobko, Roth, & Bobko, 2001; Fritz, Morris, & Richler, 2012). On a similar note, a common practice within our sample of studies was to split the items on *Gf* outcome measures in half in order to have a comparable pretest and posttest version. While this is effective in reducing test–retest practice effects, an unfortunate consequence is a reduction in measurement reliability (Jaeggi et al., 2014), which also causes downward biases in ES due to an increase in error variance that weakens the strength of its correlations (Bobko et al., 2001).

Taken together, we expect that the results reported in this meta-analysis represent a low-end estimate of the true extent of improvement that *n*-back training can have on measures of *Gf*. Moreover, our moderator analyses (described below) suggest several possible parameters that could be optimized in

Table 3 Multiple regression of effect size on control condition

Regressor	Model 1			Model 2			Model 3			Model 4		
	B	*SE*	*p*	*B*	*SE*	*p*	*B*	*SE*	*p*	*B*	*SE*	*p*
Baseline differences (covariate)	.26	.17	.13	.28	.18	.11	.30	.18	.09	.31	.18	.08
Passive control	.31	.14	.03	.23	.15	.13	.14	.17	.42	.09	.18	.53
Remuneration				−.05	.04	.22				−.04	.04	.31
International							.27	.16	.09	.25	.16	.13

Note. Multiple regression of effect size on control condition, controlling for remuneration and geographic region. Baseline differences, defined as the Hedges *g* ES between treatment and control groups at pretest, were covaried out. Passive control and international were coded as dummy variables, with active control and U.S. omitted as reference variables. Units of remuneration are in hundreds of dollars.

order to design more efficacious interventions that might have more substantive impacts on intellectual and societal functioning. In the words of Carl Sagan, extraordinary claims require extraordinary evidence, and this skepticism has rightly been applied to *n*-back training's effect on improving *Gf*. Although claims to date are ordinarily based on single studies, each with various methodological limitations, this meta-analysis of 20 different studies, 30 different treatment groups, and 24 different control groups, with data collected from 98 different *Gf* outcome measures, is a first step toward providing this extraordinary evidence.

Moderator analyses

The most striking moderator of the observed transfer effect is geographic region. International studies tend to find more transfer than U.S. studies. There is a substantial body of literature available on the effects of culture on cognition (cf. Muggleton & Banissy, 2014). These effects may contribute to differences not only between international and U.S. research participants, but also between methodological practices of researchers. However, due to the plurality of cultures (mostly European and American) represented in our meta-analysis, the precise reasons for the observed regional differences are not immediately clear. There is no difference in baseline ES between international and U.S. studies (International vs. U.S.: $g = 0.03$ vs. $g = -0.03$; $p = .73$), thereby ruling out preexisting differences between cultures or differences in participant characteristics due to the highly selective universities in which certain U.S. studies were conducted. One hypothesis, however, based on our own experiences with both U.S. and international populations, is that the former may be generally less compliant, a crucial factor in an intensive training program such as the one investigated here. To quantify this, we analyzed attrition rates reported in U.S. and international studies and found significant differences (U.S., 21%; International, 5%), $t(34) = 2.0$, $p = .05$, $d = 0.52$, across both treatment and control groups, suggesting cultural differences in attrition that may be reflective of issues surrounding general compliance and investment in research. These differences in compliance, as well as more general differences between the methodologies of international and U.S. studies, need to be tested more systematically in future research.

Additionally, there were a couple of interesting moderators that were *not* found to influence transfer. First, there was no difference in ES between research conducted by the present study authors, S.M.J. and M.B. and research conducted by other groups (Table 1). Therefore, although experimenter bias is always a prevalent concern in research, it did not significantly alter the ES of training across studies. Second, the type of control group used in studies (active or passive) did not moderate the training effect. Although our meta-analysis did reveal significantly greater transfer in studies that used passive controls, there was no significant difference in performance between the active and passive control groups themselves (Table 1; Ctrl ES), suggesting that the observed effects are likely driven by other confounding variables within passively controlled studies other than the type of control group used. Therefore, there is no evidence to support the idea that Hawthorne effects mediate the findings of transfer in studies with passive controls.

Nevertheless, the failure to reject null hypotheses of no difference between groups does not necessarily indicate the absence of an effect. As with many null findings, larger sample sizes may eventually show significant differences. However, in our analysis of research group effects (S.M.J./M.B. group vs. other research group; Table 1), there is not even a trend in any direction ($g = 0.23$ vs. $g = 0.25$). In the analysis of control groups, however, the present direction of effects actually suggests that passive control groups could end up outperforming active control groups (passive vs. active: $g = 0.28$ vs. $g = 0.08$; Table 1), which runs *opposite* to the direction suggsted by the idea that Hawthorne or expectancy effects drive improvements in both active control and treatment groups. However, it should be noted that not all of the active control groups in our analyses were of the same type. Therefore, it is possible that certain studies employ more effective active controls than do others, and true differences may exist between these more effective active controls and passive controls that may be masked in our meta-analysis.

Although the increased transfer in passively controlled studies in our analyses seems not to be driven by control group type, large improvements were found in Tx ES (Table 1), which is calculated independently of the control group. In other words, the large gap between treatment groups and passive controls (Post ES) is not due to an underperforming of passive controls, relative to active ones, but rather to an overperforming of the treatment groups (Tx ES). We sought to explain this overperformance by controlling for geographic region and remuneration. Passive control groups were mostly enrolled in international studies (10/12), as compared with active control groups (3/12), and these passively controlled studies also remunerated participants significantly less money, on average, for study participation ($61.75 vs. $220.50), $t(14) = 2.23$, $p = .04$, $d = 0.87$. These factors, individually and together, account for a sizable portion of the passive control effect and reduce its estimated effect to the point of statistical insignificance (Table 3). The potential effects of remuneration are discussed further below. Despite this, there is still a large amount of shared variance between passively controlled and international studies that needs to be teased apart in future research. Moreover, it needs to be clarified whether the relative overperformance of passively controlled or international studies represents an inflation of true effects or whether these true effects are simply masked by certain methodology associated with actively controlled or U.S. studies.

In addition to subgroup analyses, we used meta-regression to examine continuous predictors. We have previously posited that remuneration for study participation might reduce intrinsic motivation and adversely affect training outcomes (Jaeggi et al., 2014). Our regression analysis equivocally supports this hypothesis. Although we did find a negative correlation, the effect lost significance (from $p = .05$ to $p = .22$) after trimming outliers (SOM, Table S1). Nevertheless, the slope of the effect (−.07) remained the same, suggesting that the loss of significance might be indicative more of a loss of power than of the absence of an effect, particularly given our relatively small sample size. Moreover, the effects of rewards and compensation on performance have an extensive body of literature behind them, which at times can also show mixed results (Cameron, Banko, & Pierce, 2001). Therefore, while the precise nature of the effects of remuneration for *n*-back training are still unclear, it is nevertheless an important consideration for future research to at least bear in mind, given that our present data suggest that every hundred dollars a subject is compensated reduces the ES of *Gf* transfer by .07. It would be important to verify whether this trend remains with a larger sample of studies.

We also wanted to understand the type of training parameters that promote success. Although none of our analyzed parameters (session length, number of sessions, starting *n*-back level, and rate of improvement) reached significance in our main regression analyses, the SOM contains alternate regression models with trimmed outliers that revealed marginally significant negative trends between *Gf* transfer and both starting *n*-back level ($p = .06$) and session length ($p = .06$; Table S1). The former suggests that those who start with more room to improve (i.e., lower *n*-back level on day 1) may also gain the most. With regards to the latter, it is possible that shorter sessions are viewed as more achievable by participants and, therefore, more enjoyable, which might promote training quality. Longer sessions may cause fatigue that reduces motivation for subsequent sessions. Importantly, the shortest session length in our sample was only 18.5 min, so it is not clear what would happen below this data range. We predict that the inclusion of even shorter sessions would reveal an optimal training length, before and after which transfer is reduced. It is possible that our data range represents only the descending latter half of such a parabolic function.

A note of caution is warranted here regarding these regression results: given the heavy influence of outliers (SOM, Table S1) and the relatively small sample sizes, these results should be viewed as preliminary but should present a good scaffold for future research to build upon. It would be important to see how these regression trends develop in future meta-analyses that include more samples.

Conclusions

Our work demonstrates the efficacy of several weeks of *n*-back training in improving performance on measures of *Gf*. We urge that future studies move beyond attempts to answer the simple question of whether or not there *is* transfer and, instead, seek to explore the nature and extent of how these improved test scores may reflect "true" improvements in *Gf* that can translate into practical, real-world settings. On theoretical grounds, the observed improvements are plausible, since *Gf* and *n*-back performance draw upon overlapping cognitive and neural processes, including shared demands on WM and interference resolution (Burgess et al., 2011; Buschkuehl, Hernandez-Garcia, Jaeggi, Bernard, & Jonides, 2014; von Bastian & Oberauer, 2013). Since *Gf* is a fundamental cognitive skill that underlies a wide range of life functions, even small improvements can have profound societal ramifications, particularly given the healthy young adults in our analyses, representative of society's workforce. Taken together, it is becoming very clear to us that training on WM with the goal of trying to increase *Gf* holds much promise.

Author's Notes S. M. Jaeggi and M. Buschkuehl came up with the study concept and coded data. E. Sheehan and J. Au also coded data. J. Au performed the data analyses and drafted the manuscript. G. J. Duncan

consulted regarding meta-analytical techniques. All authors contributed to the data interpretation and writing, and all authors read and approved the final manuscript.

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. doi:10.1037/0033-2909.131.1.30

Basak, C., Boot, W. R., Voss, M. W., & Kramer, A. F. (2008). Can training in a real-time strategy video game attenuate cognitive decline in older adults? *Psychology and Aging, 23*(4), 765–777.

Binet, A. (1909). *Les idées modernes sur les enfants*. Paris: Ernest Flammarion Paris.

Bobko, P., Roth, P. L., & Bobko, C. (2001). Correcting the effect size of d for range restriction and unreliability. *Organizational Research Methods, 4*(1), 46–61.

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

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to Meta-Analysis*. West Sussex: John Wiley & Sons, Ltd.

Burgess, G. C., Gray, J. R., Conway, A. R., & Braver, T. S. (2011). Neural mechanisms of interference control underlie the relationship between fluid intelligence and working memory span. *Journal of Experimental Psychology. General, 140*(4), 674–692. doi:10.1037/a0024695

Buschkuehl, M., & Jaeggi, S. M. (2010). Improving intelligence: a literature review. *Swiss Medical Weekly, 140*(19–20), 266–272.

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

Cameron, J., Banko, K. M., & Pierce, W. D. (2001). Pervasive negative effects of rewards on intrinsic motivation: The myth continues. *Behavior Analyst, 24*(1), 1–44.

Caruso, D., Taylor, J., & Detterman, D. (1982). Intelligence research and intelligent policy. In D. Detterman & R. J. Sternberg (Eds.), *How and how much can intelligence be increased*. Norwood: Lawrence Erlbaum Associates.

Dahlin, E., Backman, L., Neely, A. S., & Nyberg, L. (2009). Training of the executive component of working memory: subcortical areas mediate transfer effects. *Restorative Neurology and Neuroscience, 27*(5), 405–419. doi:10.3233/RNN-2009-0492

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*(6), 627–668. discussion 692–700.

Detterman, D. K. (1993). The case for prosecution: Transfer as an epiphenomenon. In D. K. D. R. J. Sternberg (Ed.), *Transfer on trial: Intelligence, cognition, and instruction*. Norwood: Ablex Publishing Corporation.

Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: the IQ paradox resolved. *Psychology Review, 108*(2), 346–369.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*(6), 1087–1101. doi:10.1037/0022-3514.92.6.1087

Dunst, C. J., Hamby, D. W., & Trivette, C. M. (2004). Guidelines for Calculating Effect Sizes for Practice-Based Research Syntheses. *Centerscope, 3*(1), 1–10.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*(7109), 629–634.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*(1), 2-18. doi:10.1037/A0026092

Galton, F. (1892). *Hereditary genius: An inquiry into its laws and consequences*. London: Macmillan and Co.

Gottfredson, L. S. (1997). Why g matters: The complexity of everyday life. *Intelligence, 24*(1), 79–132. doi:10.1016/S0160-2896(97)90014-3

Gray, J. R., & Thompson, P. M. (2004). Neurobiology of intelligence: science and ethics. *Nature Reviews Neuroscience, 5*(6), 471–482. doi:10.1038/nrn1405

Harrison, T. L., Shipstead, Z., Hicks, K. L., Hambrick, D. Z., Redick, T. S., & Engle, R. W. (2013). Working memory training may increase working memory capacity but not fluid intelligence. *Psychological Science, 24*(12), 2409–2419. doi:10.1177/0956797613492984

Heinzel, S., Schulte, S., Onken, J., Duong, Q. L., Riemer, T. G., Heinz, A., & Rapp, M. A. (2014). Working memory training improvements and gains in non-trained cognitive tasks in young and older adults. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition, 21*(2), 146–173. doi:10.1080/13825585.2013.790338

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ, 327*(7414), 557–560. doi:10.1136/bmj.327.7414.557

Higgins JPT, Green S (editors). (2011). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.cochrane-handbook.org

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America, 105*(19), 6829–6833. doi:10.1073/pnas.0801268105

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

Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (2014). The role of individual differences in cognitive training and transfer. *Memory and Cognition, 42*(3), 464–480

Kane, M. J., Hambrick, D. Z., & Conway, A. R. (2005). Working memory capacity and fluid intelligence are strongly related constructs: comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131*(1), 66–71.

Ketay, S., Aron, A., & Hedden, T. (2009). Culture and attention: evidence from brain and behavior. *Progress in Brain Research, 178,* 79–92. doi:10.1016/S0079-6123(09)17806-8

Kundu, B., Sutterer, D. W., Emrich, S. M., & Postle, B. R. (2013). Strengthened effective connectivity underlies transfer of working memory training to tests of short-term memory and attention. *Journal of Neuroscience, 33*(20), 8705–8715. doi:10.1523/JNEUROSCI.5565-12.2013

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764–766. doi:10.1016/j.jesp.2013.03.013

Muggleton, N. G., & Banissy, M. J. (2014). Culture and cognition. *Cognitive Neuroscience, 5*(1), 1–2. doi:10.1080/17588928.2014.885781

Nisbett, R. E., & Norenzayan, A. (2002). Culture and Cognition. In D. L. Medin (Ed.), *Stevens' Handbook of Experimental Psychology, Third*

Edition. Vol 3: Memory and Cognitive Processes (pp. 561–598). Wiley Online Library: John Wiley & Sons, Inc.

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: new findings and theoretical developments. *American Psychologist, 67*(2), 130–159. doi:10.1037/a0026699

Oberauer, K., Schulze, R., Wilhelm, O., & Suss, H. M. (2005). Working memory and intelligence–their correlation and their relation: comment on Ackerman, Beier, and Boyle. *Psychological Bulletin, 131*(1), 61–65. doi:10.1037/0033-2909.131.1.61. author reply 72–65.

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. doi:10.1016/j.neuropsychologia.2013.08.012

Orwin, R. G. (1983). A fail safe *N* for effect size in meta-analysis. *Journal for Educational Statistics, 8*(2), 157–159.

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., & Engle, R. W. (2013). No evidence of intelligence improvement after working memory training: a randomized, placebo-controlled study. *Journal of Experimental Psychology. General, 142*(2), 359–379. doi:10.1037/a0029082

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. London: Sage.

Rudebeck, S. R., Bor, D., Ormond, A., O'Reilly, J. X., & Lee, A. C. (2012). A potential spatial working memory training task to improve both episodic memory and fluid intelligence. *PLoS One, 7*(11), e50431. doi:10.1371/journal.pone.0050431

Salminen, T., Strobach, T., & Schubert, T. (2012). On the impacts of working memory training on executive functioning. *Frontiers in Human Neuroscience, 6,* 166. doi:10.3389/fnhum.2012.00166

Salomon, G., & Perkins, D. N. (1989). Rocky Roads to Transfer - Rethinking Mechanisms of a Neglected Phenomenon. *Educational Psychologist, 24*(2), 113–142. doi:10.1207/s15326985ep2402_1

Schwarb, H. (2012). *Optimized Cogntive Training: Investigating the Limits of Brain Training on Generalized Cognitive Function (Doctoral Dissertation).* (Doctoral Dissertation), Georgia Institute of Technology.

Schweizer, S., Hampshire, A., & Dalgleish, T. (2011). Extending brain-training to the affective domain: increasing cognitive and affective executive control through emotional working memory training. *PLoS One, 6*(9), e24372. doi:10.1371/journal.pone.0024372

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

Stankov, L. (1986). Kvashchev Experiment - Can We Boost Intelligence. *Intelligence, 10*(3), 209–230. doi:10.1016/0160-2896(86)90016-4

Stepankova, H., Lukavsky, J., Buschkuehl, M., Kopecek, M., Ripova, D., & Jaeggi, S. M. (2014). The malleability of working memory and visuospatial skills: a randomized controlled study in older adults. *Developmental Psychology, 50*(4), 1049–1059. doi:10.1037/a0034913

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

Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology, 53*(11), 1119–1129.

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

Tomic, W., & Klauer, K. J. (1996). On the effects of training inductive reasoning: How far does it transfer and how long do the effects persist? *European Journal of Psychology of Education, 11*(3), 283–299.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine, 37*(5), 360–363.

von Bastian, C. C., & Oberauer, K. (2013). Effects and mechanisms of working memory training: a review. *Psychological Research*. doi:10.1007/s00426-013-0524-6

Wiley, J., Jarosz, A. F., Cushen, P. J., & Colflesh, G. J. H. (2011). New Rule Use Drives the Relation Between Working Memory Capacity and Raven's Advanced Progressive Matrices. *Journal of Experimental Psychology: Learning Memory and Cognition, 37*(1), 256–263. doi:10.1037/A0021613