Psychon Bull Rev (2016) 23:331–337
DOI 10.3758/s13423-015-0967-4

BRIEF REPORT

There is no convincing evidence that working memory training is NOT effective: A reply to Melby-Lervåg and Hulme (2015)

Jacky Au[1,2] · **Martin Buschkuehl**[2] · **Greg J. Duncan**[3] · **Susanne M. Jaeggi**[1,3]

Published online: 30 October 2015

Abstract Our recent meta-analysis concluded that training on working memory can improve performance on tests of fluid intelligence (Au et al., Psychon Bull Rev, 22(2), 366-377, 2015). Melby-Lervåg and Hulme (Psychon Bull Rev, doi: 10.3758/s13423-015-0862-z) challenge this conclusion on the grounds that it did not take into consideration baseline differences on a by-study level and that the effects were primarily driven by purportedly less rigorous studies that did not include active control groups. Their re-analysis shows that accounting for baseline differences produces a statistically significant, but considerably smaller, overall effect size (g = 0.13 vs g = 0.24 in Au et al.), which loses significance after excluding studies without active controls. The present report demonstrates that evidence of impact variation by the active/passive nature of control groups is ambiguous and also reveals important discrepancies between Melby-Lervåg and Hulme's analysis and our original meta-analysis in terms of the coding and organization of data that account for the discrepant effect sizes. We demonstrate that there is in fact no evidence that the type of control group *per se* moderates the effects of working memory training on measures of fluid intelligence and reaffirm the original conclusions in Au et al., which are robust to multiple methods of calculating effect size, including the one proposed by Melby-Lervåg and Hulme.

Keywords n-back · Cognitive training · Transfer · Plasticity · Meta-analysis · Fluid intelligence

Introduction

The prospect of expanding the limits of human information processing is inherently appealing. So it is not surprising that interventions such as working memory (WM) training purporting to do exactly that are met with both enthusiasm and skepticism. Our recent meta-analysis (Au et al., 2015) showing that a particular form of WM training (n-back training) can improve performance on tests of fluid intelligence (Gf) produced both of these reactions, which one way or another, have served to expand and illuminate our work (Beatty & Vartanian, 2015; Bogg & Lasecki, 2014; Deveau, Jaeggi, Zordan, Phung, & Seitz, 2014; Dougherty, Hamovitz, & Tidwell, 2015; Hughes, 2014; Karbach & Verhaeghen, 2014; Konen & Karbach, 2015; Melby-Lervåg & Hulme, 2015; Oberauer, 2015).

In an effort to continue a productive discourse on this topic, we would like to correct some misrepresentations of our meta-analysis in the recent critique by Melby-Lervåg and Hulme (2015) and also present new results from our own reanalysis of the data. Their critique concluded that the meta-analytic effect size (ES) of WM training on improving Gf was g = 0.13 (SE: 0.04), barely half the size of our previously reported ES of g = 0.24 (SE: 0.07). Though their reported ES still remained significantly different from zero, their moderator analysis suggested the effects were primarily driven by studies using passive control groups (i.e., with participants who simply took Gf tests at two time points with no intervention in-

Electronic supplementary material The online version of this article (doi:10.3758/s13423-015-0967-4) contains supplementary material, which is available to authorized users.

✉ Jacky Au
jwau@uci.edu

[1] Department of Cognitive Sciences, University of California, Irvine, CA 92697, USA

[2] MIND Research Institute, Irvine, CA, USA

[3] School of Education, University of California, Irvine, CA, USA

between). When looking only at studies with active controls (who participated in an unrelated intervention between test periods), they reported a non-significant ES of g = 0.09 (SE: 0.08) and argued that the effects of WM training in improving Gf are driven solely by placebo or Hawthorne effects (c.f., Adair, 1984), which are controlled for in studies with active but not passive controls. This criticism of our meta-analysis is not new (c.f., Dougherty et al. (2015)), and was addressed in our original paper (Au et al., 2015). Although overlooked or mischaracterized in the critiques, our original analysis included several different methods of calculating ES which, taken together, provide a more nuanced picture that do not support Hawthorne artifacts driving our conclusions.

We address this and other key issues raised by Melby-Lervåg and Hulme (ML&H). First, we disambiguate the interpretation of control group effects and argue that there is in fact no convincing evidence that n-back training effects are driven primarily by placebo or Hawthorne artifacts. Second, we review the differences between their method of ES calculation and ours, and contend that our originally reported ES of g = 0.24 is robust to either calculation method. Third, we point out various meta-analytic decisions that differed between our analyses that contribute to the discrepant ES estimates. Finally, we rebut their argument questioning the exhaustiveness of our search criteria for inclusion of studies and point out several other misrepresentations.

The issue of control groups

The methodological choice to use active or passive control groups in cognitive training studies is neither trivial nor underappreciated (c.f., Boot, Simons, Stothart, & Stutts, 2013; Rebok, 2015; Redick et al., 2013; Shipstead, Redick, & Engle, 2012; Willis, 2001). The concern with passive controls is that they do not eliminate possible placebo and Hawthorne effects that may arise in treatment groups due to expectation of improvement (Boot et al., 2013). Although plausible, evidence of a significant moderation effect of passively controlled studies, as found by both ML&H and ourselves, provides inconclusive support of placebo/Hawthorne effects. The problem is that meta-analyses, even of impacts drawn from individual studies that use random assignment, are by nature correlational in the sense that the individual studies themselves have not been randomly assigned to employ passive vs. active control groups (or to other correlated study characteristics). In fact, studies that use passive controls differ from studies that do not in a number of other important ways that can also influence the results. For example, in our original analysis, we demonstrated that passively controlled studies also tend to be conducted outside of the USA and to offer less remuneration for participation (c.f., Au et al., 2015 for theoretical motivations). Therefore, it cannot be ruled out that either of these factors (or perhaps other unmeasured ones) rather than the passive/active status of the control groups, actually cause the differential Gf transfer estimated in our sample of studies.

Although meta-analyses cannot generally disentangle the separate contributions of correlated moderators, we did argue in our original article that control groups *per se* were unlikely to be a moderator. We now illustrate this point in Fig. 1, which contrasts hypothetical data showing how Hawthorne effects *should* look (Fig. 1a) with how our meta-analytic data *actually* look (Fig. 1b). Although we observed significantly higher impacts in passively controlled studies compared to actively controlled studies, we also demonstrated that these higher impacts were not related to underperformance of passive control groups (as predicted by the Hawthorne hypothesis; Fig. 1A), but rather by overperformance of the treatment groups within passively-controlled studies (as measured by within-group changes from pretest to post-test; Fig. 1B). The reasons for this are unclear, particularly considering that treatment groups within passively controlled studies also outperform treatment groups within actively controlled studies, despite receiving the same intervention. Whatever the case, these within-group treatment effects are independent of control group performance and therefore must be the result of some other variable(s) correlated with control group type.

Furthermore, Fig. 1B clearly shows no evidence to support the Hawthorne hypothesis that active controls outperform passive controls in our data. Though there are no significant differences between them, it is noteworthy that the control group patterns actually run in the opposite direction, with passive controls performing *better* than active controls (g = 0.28 vs. g = 0.08; see Fig. 1B for ES description). ML&H seem to have misunderstood these data in their critique, contending that "the pattern [we] report (a larger effect of WM training in studies with untreated controls compared to studies with treated controls) is exactly what is expected if expectancy effects are operating to facilitate performance" (p.3). However, ML&H mistake within-group ES's of passive and active control groups across studies for between-group treatment-control ES's within actively or passively controlled studies, a subtle but important distinction. In other words, the ES's in question reflect gain scores of passive and active control groups, respectively, and not a summary ES of the treatment/control comparison. Moreover, the magnitude of this ES difference in favor of passive controls renders it difficult to argue that Hawthorne effects are being masked by a power issue (Melby-Lervåg & Hulme, 2015).

Nevertheless, interpretations of control group performance across different studies remain problematic due to methodological differences across studies. Therefore, we now present additional analyses based on the four studies in our sample that used both active and passive control groups (Fig. 2). Our results across five group comparisons (g = −0.02, SE: 0.15) reaffirm the notion that, even within the same study, there is no

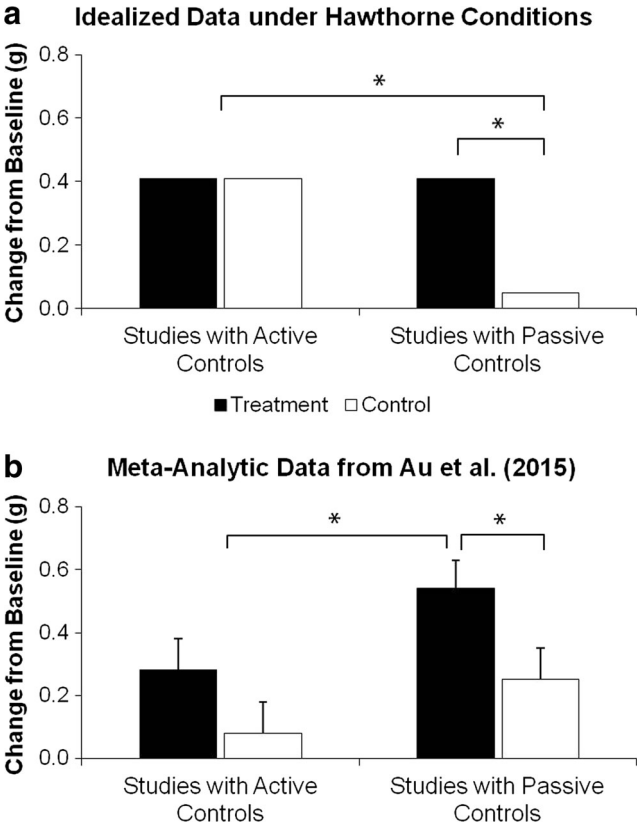

Fig. 1 (**A**) Fictional plot of idealized data assuming the effects of n-back training are driven *solely* by Hawthorne effects. Treatment groups (in both actively and passively controlled studies) improve identically to active control groups, which all perform significantly better than passive controls. (**B**) Actual meta-analytic data from the extant literature do *not* reflect the expected pattern of results shown in Fig. 1A. Effect sizes (g) in this figure are calculated as within-group standardized changes from baseline: $\frac{Post-Pre}{SD_{Pooled}}$ * $p < 0.05$

evidence of a performance difference on Gf test performance between active and passive control groups, and therefore no evidence that the effects of n-back training can be explained by Hawthorne effects. Of note, an additional relevant study (Burki, Ludwig, Chicherio, & de Ribaupierre, 2014) has come out since publication of our original meta-analysis which, although not included here since it was not in our original meta-analytic sample, also reports no differences between their active and passive control groups. Therefore, in the end, although ML&H report different numbers than we did (i.e., a smaller ES), the qualitative interpretation remains the same (i.e., a significant pooled ES driven primarily by unknown factors in studies that choose to use passive controls).

The issue of effect size calculation

We now extend our argument by reviewing the contrasting ES calculations used by ML&H in their critique and in our previous work and conclude that even the quantitative interpretation remains the same. ML&H state that the "arguably most serious problem" (p. 2) in our analysis is our use of Post ES, which takes the standardized mean difference between groups at post-test. Post ES is based on the assumption that pretest scores are homogenous between groups and therefore does not account for baseline differences. Although Post ES's are endorsed by several experts in the meta-analysis field (e.g., Dunst, Hamby, & Trivette, 2004; Higgins & Green, 2011) and although we demonstrated zero baseline differences in our sample as a whole (g = 0.003, SE = .08), ML&H argued that it is important to account for pretest differences on an individual study level. Therefore they reanalyzed our sample of studies using the standardized mean difference between gain scores of treatment and control groups, standardized by the pooled standard deviation of pretest scores (Gain Score ES; M. Melby-Lervåg, Personal Communication, 3 July, 2015).

We agree that their position makes theoretical sense, but conclude after re-analyzing the data that it makes no difference in practice (see Fig. 3). Using their method of calculating Gain Score ES (c.f., Morris, 2008) on our original dataset, we obtain g = 0.239 (SE: 0.08), which is virtually identical to the g = 0.241 (SE: 0.07) Post ES reported originally (Au et al., 2015). Therefore, we submit that the different estimates between our and ML&H's analysis (g = 0.24 vs. g = 0.13) do not depend on the type of ES used, but rather on the various (and sometimes subjective) meta-analytic decisions that take place during the coding and organization of data. We now discuss

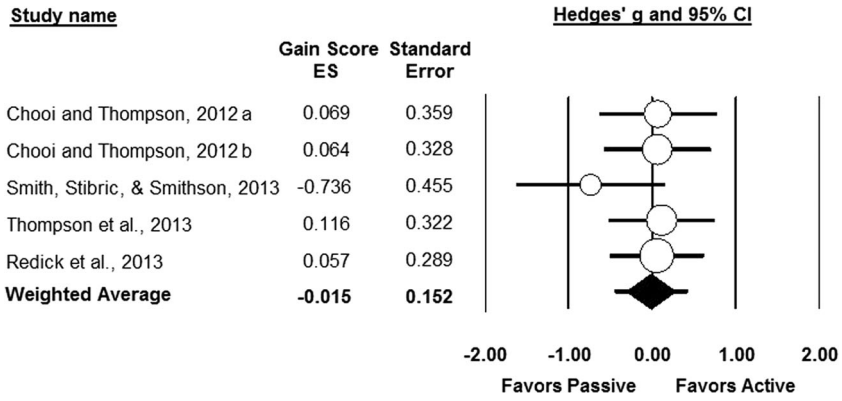

Fig. 2 Funnel plot of standardized mean differences (Hedges' g) between gain scores of active and passive control groups. We find no difference in performance on tests of fluid intelligence between active and passive control groups. Note that Chooi and Thompson (2012a,b) refer to two independent group comparisons reported in the same article

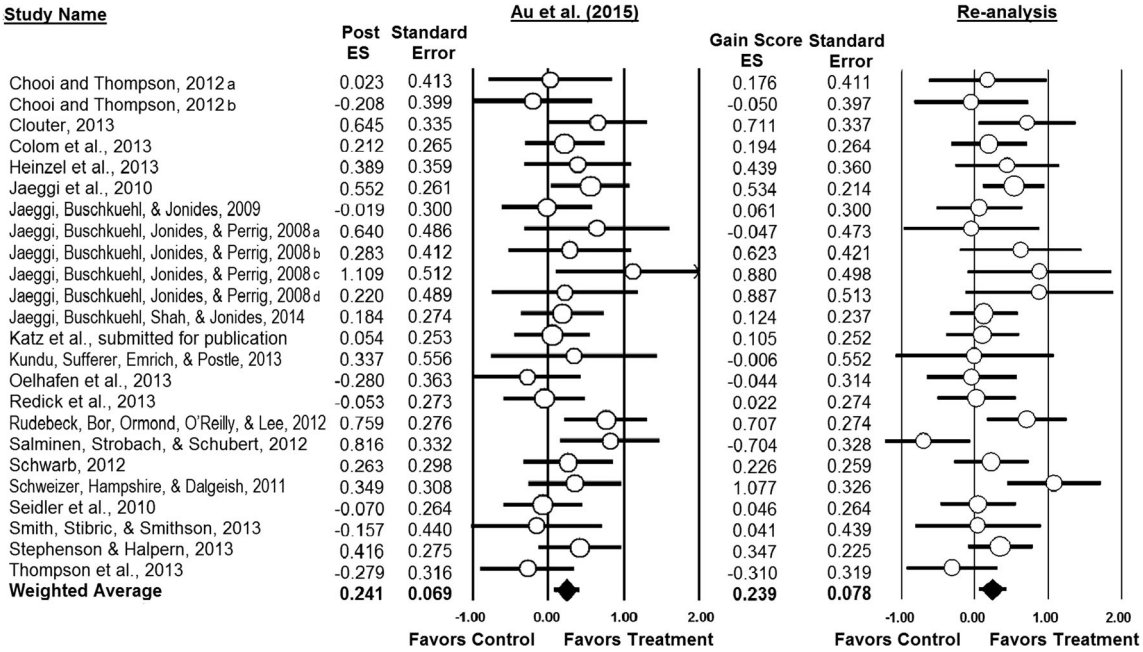

Fig. 3 A comparison of Post ES, as used in Au et al. (2015) on the left side, and Gain Score ES, as recommended by Melby-Lervåg and Hulme (2015) on the right side. When calculated from the same dataset (Au et al., 2015), the overall weighted average is virtually identical in both instances. Forest plots show Hedges' g ± 95 % confidence intervals. Post ES: $\frac{Post_{Treatment} - Post_{Control}}{SD_{Pooled}}$, Gain Score ES: $\frac{(Post_{Treatment} - Pre_{Treatment}) - (Post_{Control} - Pre_{Control})}{SD_{Pooled\ Pre}}$ Note that letters after the year of publication refer to different group comparisons reported in the same article

several important decisions that account for much of the discrepancy.

The issue of meta-analytic decisions

The most serious discrepancy between our analyses is the number of independent comparisons: 39 in ML&H and 24 in our original paper. Although ML&H expressed uncertainty as to how we arrived at only 24 comparisons, our original report characterized independent comparisons as "an independent set of ESs such that each treatment or control group was never represented more than once in the overall analysis" (p.370). This meant that "ES's 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" (p. 370). ML&H, on the other hand, treated each treatment-control comparison in a single study as independent, generating a separate ES for each, even when they are compared with the same control group. Although this method would appear to increase the power of their analysis relative to ours, giving them 39 comparisons to our 24, the lack of independence in their comparisons invalidates tests of statistical significance that do not account for this clustering (c.f., Borenstein, Hedges, Higgins, & Rothstein, 2009). In addition, it leads to over-representation of certain control groups that can magnify any peculiarities or idiosyncrasies uniquely endemic to any one of them. Furthermore, ML&H decided to code data from all control groups within a study, whereas we only chose the active over the passive if both existed in the same study (p. 368). While both approaches are legitimate, it becomes problematic for the same reasons just described above when ML&H compare the same treatment group to multiple control groups, and calculate separate ES's for each comparison.

An illustration of this non-independence problem is found in Stephenson and Halpern (2013). This study had three intervention groups (dual, single, and auditory n-back), and two control groups (active and passive). The three intervention groups all showed a fairly similar level of improvement, as did the active, but not the passive, control group. However, the active control group actually trained on an adaptive spatial matrix task that required holding a spatial sequence in WM. Since the purpose of our meta-analysis was to evaluate the efficacy of WM training, we did not consider this to be an appropriate control and excluded it from our analysis, which we clearly described in our Methods section (p. 368). However, not only did ML&H decide to include this control group, they included it three times, generating a separate ES for each comparison against the three n-back interventions. We, on the other hand, excluded the active control, and averaged the three treatment groups together into one and compared the aggregate treatment group to the passive control

group in order to generate only one independent comparison from this study.

A similar issue arises in Smith et al. (2013), which employed both an active and a passive control group, as well as a second treatment group that received video game training. The aim of that study was to test the efficacy of n-back training and video game training in improving Gf compared with both an active and a passive control. However, ML&H also treated the video game treatment group as a third control group, resulting in three treatment-control comparisons. Moreover, this n-back training group improved minimally at post-test (d = 0.10), but sharply at a 1-week follow-up (d = 0.58), most likely reflective of imprecise estimation due to the small sample size (n = 10). Nevertheless, the imprecisely estimated null ES at post-test is reflected not just once in ML&H's analysis, but three times, including an improper comparison against another treatment group also hypothesized by the original authors to improve Gf.

A final issue is the classification of Gf tasks. ML&H chose to redefine the analysis to measure "nonverbal reasoning" rather than Gf, *per se*. Though the two overlap substantially, Gf is not restricted purely to nonverbal domains (Ackerman, Beier, & Boyle, 2005; Kane et al., 2004). Accordingly, ML&H took issue with our inclusion of Reading Comprehension as a Gf outcome measure in their critique, and there were several others of our outcomes which they excluded as well. There is no consensus as to which tests constitute measures of Gf and which do not. Though we chose to define our measures based on guidelines set forth by Ackerman et al. (2005) and Gray and Thompson (2004), other selection criteria such as the ones outlined by ML&H are of course legitimate. But, we point out this difference as a potentially important distinction between our two analyses, and refer readers to the Supplementary materials of our original analysis (Au et al., 2015) for a complete list of all our outcome measures used.

Other relevant issues

Aside from the major issues concerning ES calculations and control groups, there were several other misrepresentations and errors we would like to correct. First of all, ML&H question the comprehensiveness of our inclusion criteria using Pubmed and Google Scholar, pointing out that they found three additional articles not included in our analysis by searching PsycInfo and ERIC (Anguera et al., 2012; Colom et al., 2010; Nussbaumer, Grabner, Schneider, & Stern, 2013). However, relying in large part on Google Scholar as one of the most comprehensive databases available and being active researchers in the field of WM training, we were well aware of these three studies. We excluded Anguera et al. (2012), on which two of us are authors, because the Gf data are identical to those reported in Seidler et al. (2010), which was already included in our analysis. Colom et al. (2010) did not include any form of n-back training whatsoever, and clearly did not meet our inclusion criteria. And Nussbaumer et al. (2013) was excluded due to missing data relevant to ES calculations[1] (as indicated in our flow chart in Fig. 1 of Au et al., 2015). The fact that ML&H included the Colom et al. study in their re-analysis is both erroneous and consequential, because the study generated a negative (g = −0.15) ES that was Heavily weighted owing to its large sample size (N = 173), which is nearly an order of magnitude larger than the average size of other studies in the sample (average N ± SD = 19.96 ± 8.13).

Additionally, ML&H state that we did not include information about the start and end dates in our search. On the contrary, we clearly stated (p. 370) that all the studies in our analysis were completed between 2008 and 2013 (though some appeared in print later than 2013). Also, ML&H point out that seven of our studies were not listed in the bibliography. This was unfortunately an error during the copy-editing process. We had originally included in our Supplementary materials a complete list of references, which was not faithfully transferred online. We thank them for pointing this out and have included this now in the Supplementary materials of the present work.

Finally, ML&H describe our meta-analysis several times in their critique as being "less than transparent." However, we maintain that all of our procedures and meta-analytic decisions were clearly described in our paper, including our search criteria and how we arrived at 24 independent comparisons, both of which issues were specifically criticized by ML&H as being non-transparent. Nevertheless, we agree that providing information about our individual study characteristics, including coded measures and associated ESs, would facilitate replication. We now include this information with this current report in the Supplementary online materials.

Conclusion

Our original article concluded that "it is becoming very clear to us that training on WM with the goal of trying to increase Gf holds much promise." (p. 375). Despite ML&H's critique, we still stand by this statement. Having addressed their criticisms, we find that neither the qualitative nor quantitative

[1] The authors were contacted for this information but did not provide it. We note that these data are now available through Melby-Lervåg & Hulme (2015), but are not included here in order to maintain a valid comparison to our original meta-analysis. The same goes for Burki et al. (2014), which fits our inclusion criteria and was included in Melby-Lervåg & Hulme (2015), but was published after our original article.

interpretations of our original work change. There still seems to be an overall small, but significant ES of n-back training on improving Gf test performance. These effects cannot easily be explained as Hawthorne effects or artifacts of control group type. We continue to urge that the next steps of research in this field should seek to isolate the conditions under which these effects can most reliably manifest, and to seek demonstrations of practical, real-world gains in activities requiring Gf.

ML&H have criticized that even if the ES of n-back training could be taken at face value, the effects may still be too small to be of practical significance. We concede that this is possible, especially since it is unclear to what extent an ES of g = 0.24 on laboratory tests of Gf translates to real-world gains in actual intelligence. However, any true improvement in intelligence, no matter how small, is of interest from a basic science perspective if not a translational one. Any convincing proof of concept would enable a fruitful avenue of research into isolating and augmenting the source of the effect. Furthermore, it is promising that this meta-analytic effect was demonstrated in young, healthy adults who were mostly university students already at or near the peak of their cognitive abilities, leaving open the question of whether cognitively sub-optimal populations might benefit even more (c.f., Weicker, Villringer, & Thöne-Otto, 2015) .

Author note This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1321846. JA and MB are employed at the MIND Research Institute, whose interest is related to this work. SMJ has an indirect financial interest in MIND Research Institute.

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

Adair, J. G. (1984). The Hawthorne Effect - a Reconsideration of the Methodological Artifact. *Journal of Applied Psychology, 69*(2), 334–345. doi:10.1037/0021-9010.69.2.334

Anguera, J. A., Bernard, J. A., Jaeggi, S. M., Buschkuehl, M., Benson, B. L., Jennett, S., … Seidler, R. D. (2012). The effects of working memory resource depletion and training on sensorimotor adaptation. *Behavioural Brain Research, 228*(1), 107–115. doi:10.1016/j.bbr.2011.11.040

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

Beatty, E. L., & Vartanian, O. (2015). The prospects of working memory training for improving deductive reasoning. *Front Hum Neurosci, 9*. doi:10.3389/Fnhum.2015.00056

Bogg, T., & Lasecki, L. (2014). Reliable gains? Evidence for substantially underpowered designs in studies of working memory training transfer to fluid intelligence. *Frontiers in Psychology, 5,* 1589. doi:10.3389/fpsyg.2014.01589

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. doi:10.1177/1745691613491271

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Chatper 25: Multiple comparisons within a study introduction to meta-analysis*. West Sussex, UK: John Wiley & Sons, Ltd.

Burki, C. N., Ludwig, C., Chicherio, C., & de Ribaupierre, A. (2014). Individual differences in cognitive plasticity: An investigation of training curves in younger and older adults. *Psychological Research, 78*(6), 821–835. doi:10.1007/s00426-014-0559-3

Colom, R., Quiroga, M. A., Shih, P. C., Martinez, K., Burgaleta, M., Martinez-Molina, A., … Ramirez, I. (2010). Improvement in working memory is not related to increased intelligence scores. *Intelligence, 38*(5), 497–505. doi:10.1016/j.intell.2010.06.008

Deveau, J., Jaeggi, S. M., Zordan, V., Phung, C., & Seitz, A. R. (2014). How to build better memory training games. *Frontiers in Systems Neuroscience, 8,* 243. doi:10.3389/fnsys.2014.00243

Dougherty, M. R., Hamovitz, T., & Tidwell, J. W. (2015). Reevaluating the effectiveness of n-back training on transfer through the Bayesian lens: Support for the null. *Psychonomic Bulletin and Review*. doi:10.3758/s13423-015-0865-9

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.

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

Higgins, J. P. T., & Green, S. (Eds.). (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

Hughes, J. (2014). Enhancing virtues: Intelligence (Part 2). Retrieved July 24, 2015, 2015, from http://ieet.org/index.php/IEET/more/9508

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133*(2), 189–217. doi:10.1037/0096-3445.133.2.189

Karbach, J., & Verhaeghen, P. (2014). Making working memory work: A meta-analysis of executive-control and working memory training in older adults. *Psychological Science, 25*(11), 2027–2037. doi:10.1177/0956797614548725

Konen, T., & Karbach, J. (2015). The benefits of looking at intraindividual dynamics in cognitive training data. *Frontiers in Psychology, 6*. doi:10.3389/Fpsyg.2015.00615

Melby-Lervåg, M., & Hulme, C. (2015). 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 and Review*. doi:10.3758/s13423-015-0862-z

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

Nussbaumer, D., Grabner, R., Schneider, M., & Stern, E. (2013). *Limitations and chances of working memory training*. Paper presented at the Proceedings of the 35th Annual Meeting of the Cogntive Science Society, Berlin, Germany.

Oberauer, K. (2015). Training of intelligence: A question of intelligent training. Retrieved from http://www.psychonomic.org/featured-content-detail/training-of-intelligence-question-of-intelligent-t

Rebok, G. W. (2015). Selecting control groups for randomized controlled trials of behavioral interventions. In L. N. Gitlin & S. J. Czaja (Eds.), *Behavioral intervention research: Designing, testing, and implementing*. New York: Springer Publishing Company.

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

Seidler, R., Bernard, J., Buschkuehl, M., Jaeggi, S. M., Jonides, J., & Humfleet, J. (2010). *Cognitive training as an intervention to improve driving ability in the older adult*. Ann Arbor, MI: University of Michigan.

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

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

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

Weicker, J., Villringer, A., & Thöne-Otto, A. (2015). Can impaired working memory functioning be improved by training? a meta-analysis with a special focus on brain injured patients. *Neuropsychology*. doi:10.1037/neu0000227

Willis, S. L. (2001). Methodological issues in behavioral intervention research with the elderly. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (5th ed., pp. 78–108). San Diego, CA: Academic Press.