Psychonomic Bulletin & Review

# Re-evaluating the effectiveness of n-back training on transfer through the Bayesian lens: Support for the null.

| | |
|---|---|
| Journal: | *Psychonomic Bulletin & Review* |
| Manuscript ID: | PBR-Otr-14-049.R2 |
| Manuscript Type: | Other |
| Date Submitted by the Author: | 13-May-2015 |
| Complete List of Authors: | Dougherty, Michael; University of Maryland, Psychology<br>Hamovitz, Toby; Univeristy of Maryland, Psychology<br>Tidwell, Joe; Univeristy of Maryland, Psychology |
| Keywords: | working memory training, N-back, fluid intelligence, Bayes Factors |
| | |

Running head: N-back Training and Gf

# Re-evaluating the effectiveness of n-back training on transfer through the Bayesian lens: Support for the null

Michael R Dougherty

Toby Hamovitz

&

Joe W. Tidwell

Department of Psychology

University of Maryland

Contact information:

Michael Dougherty
Department of Psychology
University of Maryland
College Park, MD 20742
301-405-8423
mdougher@umd.edu

## Abstract

A recent meta-analysis by Au et al. (2015) reviewed the n-back training paradigm for working memory (WM) and evaluated whether when aggregating across existing studies, there was evidence that gains obtained for training tasks transferred to gains in fluid intelligence (Gf). Their results revealed an overall effect size of *g*=0.24 for the effect of n-back training on Gf. We re-examine the data through a Bayesian lens, to evaluate the relative strength of the evidence for the alternative versus null hypotheses, contingent on the type of control condition used. We find that studies using a non-contact (passive) control group strongly favor the alternative hypothesis that training leads to transfer, but that studies using active control groups show modest evidence in favor of the null. We discuss these findings in the context of placebo effects.

Key words: *Working memory training, N-back, placebo effects, meta-analysis, Bayes factors*

Perhaps one of the most exciting, yet controversial, areas of research within the psychological sciences concerns the effectiveness of working memory (WM) training for improving general cognitive functions. The mere possibility that core WM processes can be improved remains an enticing idea for the simple reasons that WM is central to performance on a wide range of daily activities (Engle, 2002) and because deficits in WM are associated with numerous clinical disorders (e.g., Willcut et al., 2005). While the potential implications of WM training for society are widely agreed upon, the strength of the evidence supporting the effectiveness of WM training is debatable. Although numerous studies show apparent transfer effects to measures of general cognitive abilities (Chein & Morrison, 2010; Oei & Patterson, 2013), many other studies fail to yield positive results (e.g., Rode et al., 2014; Sprenger et al., 2013). The lack of consensus across individual studies is striking, and raises many questions about the robustness of the effect, as well as how moderator variables may determine the boundary conditions under which training reliably leads to improvements in untrained general cognitive abilities.

A major problem underlying many claims of WM training effectiveness, regardless of whether significant effects obtain, is the reliance on small samples. It is for this reason that meta-analytic techniques, such as those employed by Au, Sheehan, Tsai, Duncan, Buschkuehl, and Jaeggi (2015) are necessary. In their meta-analysis of the impact of training on n-back, Au, et al. provide a compelling case for the impact of n-back training on measures of Gf. Combined across 20 studies, Au et al. revealed that there was a statistically reliable effect of n-back training on Gf transfer tasks. Although the observed effect size was small ($g$=0.24), even small improvements in core cognitive functions such as working memory and Gf could have enormous societal implications.

However, in contrast to Au et al., we *do not* agree that the data included in their meta-analysis of n-back training warrant the conclusion that “short-term cognitive training … can result in beneficial effects in important cognitive functions” (pg. 366). Although their analysis does indeed illustrate an effect of some sort, we propose that this effect is an experimental design artifact, and is consistent with a placebo effect interpretation~~.~~

In what follows, we lay out the basis for our claim, which is leveraged on reinterpreting the evidence provided by Au et al. (2015) through a Bayesian lens. Specifically, we reconsider the importance of using proper control conditions, as well as accounting for the null hypothesis as a theoretically relevant alternative. While we commend Au et al. on a rigorous meta-analysis, we contend that their analysis insufficiently address these issues. For example, while Au et al relied on well-established null hypothesis significance testing (NHST) methods for meta-analysis, two well-known limitations of the NHST framework are that it both tends to overstate evidence for the alternative hypothesis, and does not permit one to evaluate the relative probability that the null hypothesis is in fact true[1]. In the context of the WM training literature, both of these problems are especially salient because the primary issue of debate is *if* working memory training is effective at all. This implies a need to evaluate the degree to which the data support the alternative hypothesis *relative* to the null, and is most easily addressed within a Bayesian approach.

The second issue relevant to our reanalysis concerns the need to use proper control conditions. In the medical literature, the gold standard for evaluating the

[1] The limitations of NHST methods are well documented and need not be rehashed here in their entirety; readers interested in this topic are invited to read papers by Raftery (1995), Wagenmakers (2007), Rouder et al. (2009), and in particular Rouder and Morey (2011) and Rouder, Morey, and Province (2013).

effectiveness of pharmaceuticals is the double-blind placebo control study where neither the study moderator (e.g., the experimenter) nor the participant knows what condition to which he or she is assigned. The purpose of using double-blind placebo control groups is to control for potential effects due to participants' expectations, which can be induced either by direct knowledge of the intervention or by being treated differently by the researcher.[2] Unfortunately, deviation from the double-blind procedure is the norm within the cognitive training literature: We know of only a small number of studies that attempted to use a double blind placebo-control procedure (Sprenger et al., 2013, study 2; von Bastian & Eschen, in press). Most studies either use a no-contact control condition (no placebo control and often referred to as a passive control), or a single-blind placebo control (often referred to as an active control) in which the experimenter interacting with the subjects knows group assignment, but the participant is blinded (as much as possible) to whether he or she was assigned to the true intervention or a sham intervention. Even within studies using active controls, there is considerable heterogeneity on the specific nature of the control. Some use control tasks that are designed to look like the training tasks (e.g., visual attention training, Redick et al., 2013), but without the efficacious properties theoretically needed to promote improvement in WM; others use control tasks that are ostensibly different from the training tasks (e.g., knowledge training, Jaeggi, Buschkuehl, Jonides, & Shah, 2011; Jaeggi, Buschkuehl, Shah, & Jonides, 2014). The comparison of training effects relative to a properly chosen control is paramount for establishing training effectiveness, since the precise nature of the intervention cannot be

[2] This is particularly challenging for cognitive training studies because participants literally see and engage in the intervention, making it difficult to mask what condition participants believe they have been assigned to.

entirely concealed from the participant (see Boot, Simons, Stothart, & Stutts, 2013 for a recent discussion of placebo controls). Without showing effects relative to a proper control condition, or otherwise controlling for possible placebo effects, it is difficult to move forward with interpreting results from the passive control studies, let alone justify claims of effective transfer. If n-back training does indeed produce gains in fluid abilities, as claimed by Au et al., then this should hold both for studies that include passive control groups and for studies that use active control groups. Although the nature of the control tasks differ considerably across studies characterized as involving active-control groups, we assume that these studies represent more appropriate control conditions compared to studies using no-contact or passive controls.

## Preliminaries

An important component of meta-analyses entail selecting those studies that should be included. Au et al. (2015) made an excellent attempt to reduce the potential influence of publication bias, with many studies included from non-published reports. The selection of studies to be included in the analysis appears to have been thorough and fair. Two important details of the selection criteria are that Au et al. limited their analyses to studies that used a form of the N-back task as the only training task and to studies that included healthy adults aged 18-50. Thus, the conclusions we draw below do not necessarily generalize to other types of training or age groups. Critically, Au et al. (2015) included two types of studies in their analysis: Those that used passive controls and those that used active controls. This is critical because whether the study includes a passive control or an active control will dictate the degree to which the results are open to alternative interpretations, such as a placebo effect.

Au et al. presented effect sizes for 24 individual comparisons drawn from 20 papers. The aggregate weighted effect size across these 24 comparisons was 0.24. They also evaluated several possible mediators, including whether the studies used an active control (N=12) or a passive control (N=12), which yielded effect sizes of 0.06 and 0.44 respectively. Although Au et al. reported this effect as significant, they concluded that type of control group did not moderate the effect. This strikes us as an odd conclusion given that the magnitudes of these effect sizes differ considerably.[3] The question is: Do these effect sizes provide evidence for training effectiveness?

## The Bayesian Analysis of Transfer Effects

We re-analyzed the data contained in Figure 3 of Au et al. (2015) from a Bayesian perspective, which provides a more natural way of interpreting the strength of evidence. As intimated above, a feature of the Bayesian analysis is that it permits one to evaluate the likelihood of the data under both the null hypothesis of no transfer to Gf and the alternative hypothesis that N-back training transfers to Gf. Our analysis approach closely followed the methods used by Rouder and Morey (2011) and Rouder, Morey, and Province (2013) in their meta-analyses of psi. Further, we used only those effect sizes included in Figure 3 of the Au et al. paper and we retained the scheme used to categorize studies as using active or passive control.

[3] Au et al.'s conclusion was based on a comparison between the control groups for active and passive studies, not by comparing the control groups to the treatment condition. The comparison of control groups while ignoring the training groups isn't particularly informative regarding effect of training, since the effects of training can only be assessed relative to the control. In this regard, it is interesting to note that the effect size for the training condition amongst active-control studies (d=0.25) is actually numerically smaller than the effect size amongst the control participants in the passive control studies (d=0.28).

The first step of our analysis involves transforming the effect sizes presented in Figure 3 of Au et al. to their corresponding t-values using t=sqrt(1/n1+1/n2)*g, where g is the measure of effect size and n1 and n2 are the sample sizes for two independent groups used in the effect size calculations. We then computed the default Bayes Factor corresponding to each t-statistic using the ttestBF function in the BayesFactor package in R (R Core Team, 2014; Morey, Rouder, & Jamil, 2014), as well as the meta-analytic Bayes Factor using the meta.ttestBF function. For all analyses, we set the scale factor on effect size to r = 1, and used a one-sided interval, which places the mass of the prior on effects greater than zero. The one sided test is a reasonable assumption under the hypothesis that training should lead to *improvements* in Gf. Importantly, even large modifications to the prior distribution do not alter our conclusions in any substantive way, nor does using a two-sided null interval.

The Bayes Factors for each study are presented in Figure 1. The values of g, t, and sample sizes used in our analysis for each study are presented in Table 1. We have organized Figure 1 and Table 1 by study type (active versus passive control) with the individual studies in Figure 1 sorted by the magnitude of the BF. As a point of reference, it is standard to interpret magnitudes of the BF along a graded scale such that values between 1 and 3 provide weak evidence for the alternative and values between 1/3 and 1 provide weak evidence for the null; BF's between 3 (1/3) and 10 (1/10) are interpreted as 'substantial' evidence; BF's between 10 (1/10) and 30 (1/30) are interpreted as 'strong', and values over 100 (1/100) are interpreted as 'decisive' (Jeffreys, 1961)[4].

[4] Jeffrey's (1961) labeling scheme provides one set of guidelines for interpreting the magnitude of BF's. Although others have been proposed (e.g., Kass & Raftery, 1995),

**Figure 1 A**

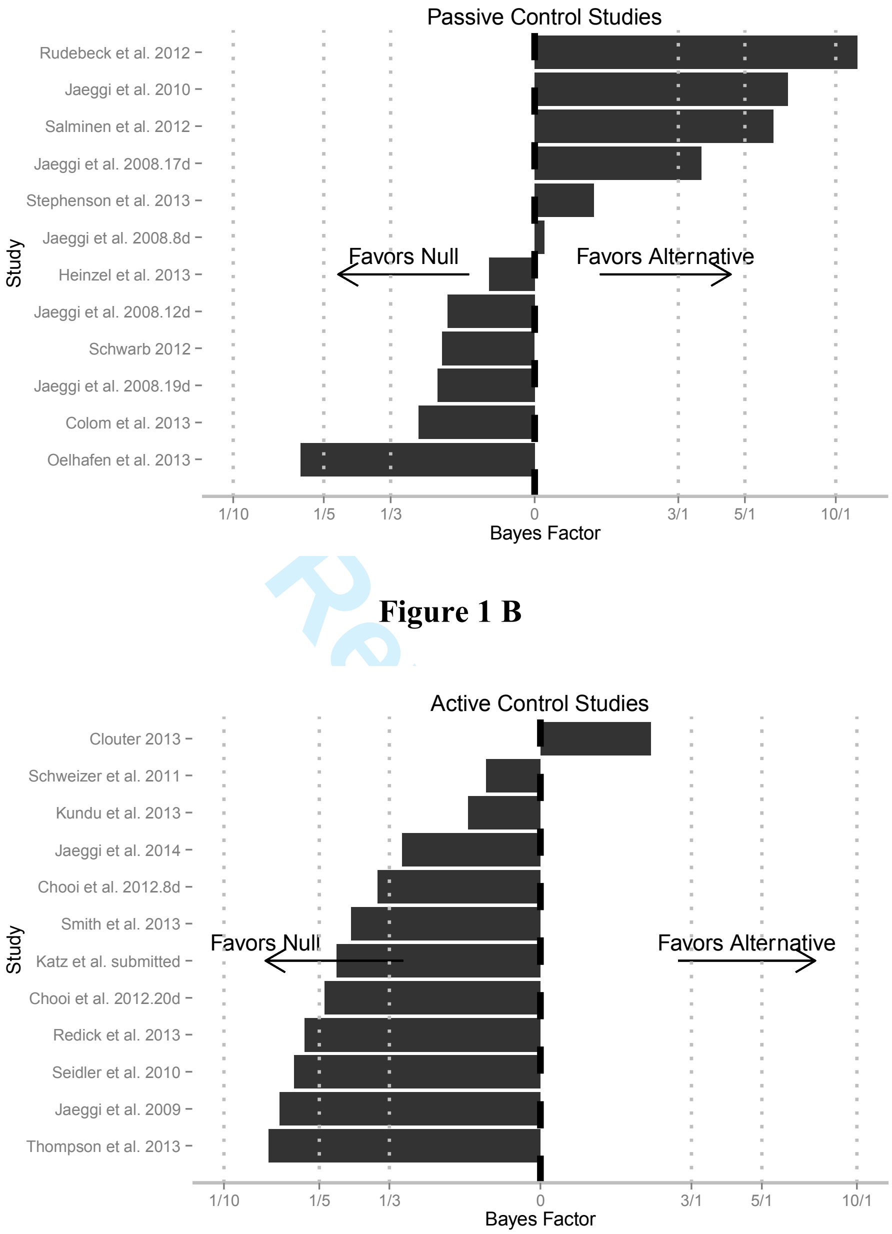


**Figure 1 B**

**Figure 1. Panel A plots the Bayes Factor for the 12 comparisons that used a passive control. Panel B plots the Bayes Factor for the 12 comparisons that used active controls. The study label includes days of training for studies that included between**

the beauty of the BF is that it can be interpreted numerically as the strength of evidence for a particular model relative to an alternative.

**groups manipulations of length of training (e.g., Jaeggi et al. 2008.8d corresponds to the condition in which participants trained for 8 days on n-back).**

**Table 1. Descriptive statistics for each study included in the meta-analysis.**

| Experiment | t | n1 | n2 | Hedge's g |
|---|---|---|---|---|
| **Passive Control Studies** | | | | |
| Rudebeck et al. 2012 | 2.813 | 27 | 28 | 0.759 |
| Jaeggi et al. 2010 | 2.602 | 46 | 43 | 0.552 |
| Salminen et al. 2012 | 2.511 | 20 | 18 | 0.816 |
| Jaeggi et al. 2008.17d | 2.218 | 8 | 8 | 1.109 |
| Stephenson et al. 2013 | 1.848 | 82 | 26 | 0.416 |
| Jaeggi et al. 2008.8d | 1.28 | 8 | 8 | 0.64 |
| Heinzel et al. 2013 | 1.065 | 15 | 15 | 0.389 |
| Schwarb 2012 | 0.872 | 22 | 22 | 0.263 |
| Colom et al. 2013 | 0.793 | 28 | 28 | 0.212 |
| Jaeggi et al. 2008.12d | 0.663 | 11 | 11 | 0.283 |
| Jaeggi et al. 2008.19d | 0.425 | 7 | 8 | 0.22 |
| *Oelhafen et al. 2013* | *-0.753* | *14* | *15* | *-0.28* |
| **Active Control Studies** | | | | |
| Clouter 2013 | 1.935 | 18 | 18 | 0.645 |
| Schweizer et al. 2011 | 1.12 | 29 | 16 | 0.349 |
| Kundu et al. 2013 | 0.859 | 13 | 13 | 0.337 |
| Jaeggi et al. 2014 | 0.773 | 51 | 27 | 0.184 |
| Katz et al. submitted | 0.2121 | 36 | 27 | 0.054 |
| Chooi et al. 2012.8d | 0.054 | 9 | 15 | 0.023 |
| *Redick et al. 2013* | *-0.192* | *24* | *29* | *-0.053* |
| *Seidler et al. 2010* | *-0.261* | *29* | *27* | *-0.07* |
| *Smith et al. 2013* | *-0.341* | *10* | *9* | *-0.157* |
| *Chooi et al. 2012.20d* | *-0.507* | *13* | *11* | *-0.208* |
| *Jaeggi et al. 2009* | *-0.6227* | *22* | *21* | *-0.19* |
| *Thompson et al. 2013* | *-0.861* | *20* | *19* | *-0.276* |

As should be evident from Figure 1 and Table 1, few of the individual studies provide particularly strong evidence for either the null or the alternative. Yet looking across the entirety of the results a curious pattern is obvious. First, 11 of the 12 effect sizes for the passive control studies are positive, whereas only 6 of the 12 effect sizes are positive for the active-control studies. Second, when these effect sizes are evaluated in terms of the Bayes Factor, the majority of the individual studies favor the null hypothesis, including 6 of the 12 passive-control studies. These individual results using the BF roughly mirror the conclusions drawn from the significance tests, though the BF illustrates that the bulk of the studies show evidence for the null. However, these individual comparisons do not capitalize on a major strength of meta-analytic techniques, which is the ability to aggregate across studies to overcome the sample size problem.

Moving on to the meta-analytic results, here the results diverge somewhat from the conclusions garnered from the individual studies. First, ignoring the type of control, the odds in favor of the alternative hypothesis is 152:1. This qualifies as 'decisive' evidence according to Jeffreys' (1961) scheme. Figure 2, which provides the BF's conditioned on the use of passive versus active control groups, paints a much different picture. While the Bayes factor for the passive control studies is a whopping 13,241:1 in favor of the *alternative*, the Bayes factor for the active control studies is a more modest 7.7:1 in favor of the *null*. As stated above, this qualifies as *substantial* evidence for the null. Note that when a two-sided test is used lieu of the one-sided test, we obtain a Bayes factor of 6000:1 (in favor of the alternative) for the passive control studies and a Bayes factor of 11:1 (in favor of the null) for the active control studies. The later constitutes *strong* evidence for the null amongst studies using proper experimental controls.

**Figure 2**

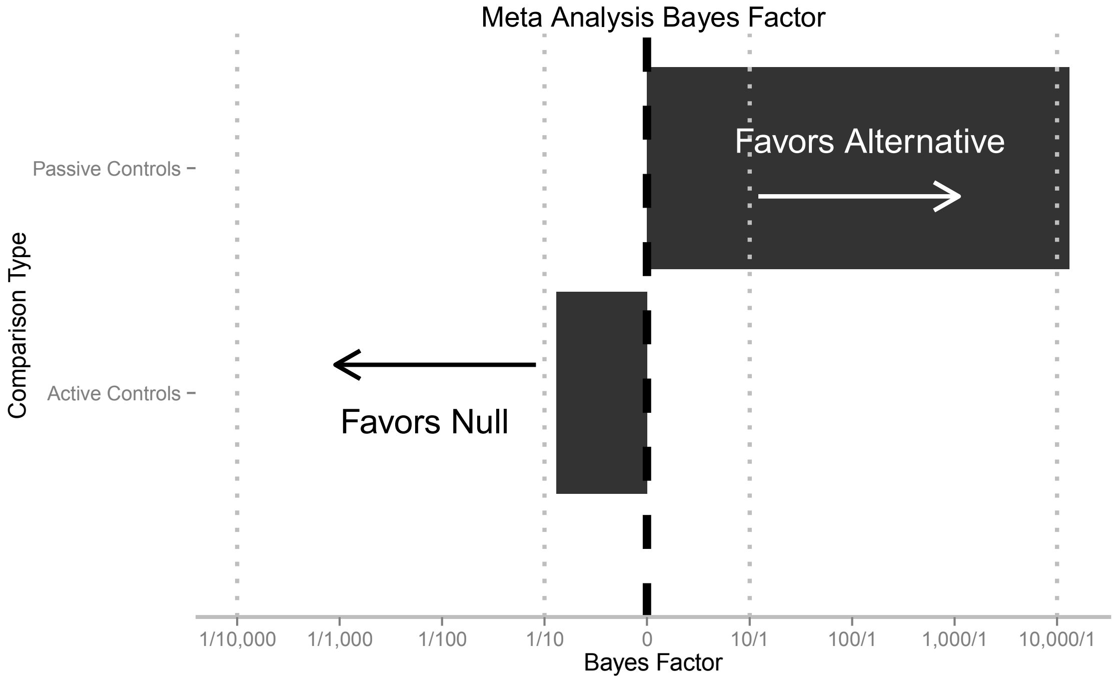


**Figure 2. Meta Analysis Bayes Factor. Compares overall BF for the two subgroups of passive and active controls assuming a one-sided interval in favor of the alternative.**

One potential objection to our BF analysis above is that we segregated the data by type of control condition, rather than modeling effects as a function of control type. This is a relevant objection because splitting the data by control-type ignores an important source of variability that can enable more precise estimates of effect sizes. Thus, we conducted a series of follow-up analyses using hierarchical Bayesian modeling in which we modeled the effect sizes as a function of Control group type (passive versus active), as well as an additive effect of both control group type and country of origin (USA versus non-USA). Au et al. (2015) identified country of origin as an important moderator variable, with studies conducted within the USA yielding a small non-significant effect size and studies conducted outside the USA resulting in a moderate significant effect size – an effect that Au et al. hypothesized could be due to differences in motivation or compliance between USA and non-USA subjects. The inclusion of country of origin in our analysis allowed us to control for a potential important source of variability that Au et al. (2015) felt was theoretically justified. As we illustrate below, inclusion of this variable in the Bayesian model reveals that the only estimated effect sizes that are different from zero are those based on non-USA passive control studies. Further, the estimated effect size for the active control studies within the USA shrink to essentially zero.

The hierarchical Bayesian analyses were repeated using three different prior distributions on the population level effect size (Table 2) to assess the sensitivity of the posterior distribution to different prior beliefs. In practical terms, we modeled what one should believe about the effect of training on transfer, given the evidence from the studies included in the meta-analysis and given whether *a priori* one either has no prior knowledge of an effect, or has knowledge corresponding to a small, medium, or large prior population effect size (i.e., these priors were set such that they favor the hypothesis that training is effective). Table 2 provides the relevant parameters and prior probability distributions for these four model variants. R code for running the hierarchical models is provided as supplemental material.

The results of the hierarchical Bayesian analyses are presented in Figures 3 through 6, which plot median estimated effect sizes (with 95% Highest Density Intervals) for both a simple meta-analytic model (intercept only model) and the models that include the effect of control type (passive versus active) and country of origin (USA versus non-

USA). The model that includes only the effect of control type is plotted in Figures 3 (individual study estimates) and 4 (aggregate effect sizes). There is a clear discrepancy between studies that include passive versus active control groups: Studies that include a passive control consistently show positive effect sizes whereas the studies that include an active control consistently obtain an effect size only marginally greater than 0, as evident by the fact that the estimated effect sizes and HDIs are nearly centered on 0. This result is consistent across different prior distributions. Strikingly, even when the prior distribution is set such that the effect of training is assumed to be large, there is still no evidence of that N-back training leads to improvements on Gf measures. While this model estimates that the median effect size amongst the active control studies is slightly above 0, this small positive effect is essentially eliminated when country of origin is added as a predictor in the model as shown in Figures 5 and 6. Importantly, the three international studies using active controls fail to yield a reliable positive effect. Further, at the aggregate level the only effect size in which the HDI does not include zero are effects based on studies conducted outside the USA that use passive control designs. Again, the conclusions drawn from modeling by control type and country of origin as predictors of effect size are consistent across different prior assumptions.

**Table 2: Parameter values for prior distribution of effect size, N(μ,σ) , and the probability that the effect size is greater than 0, *p*(g > 0)**

| Prior on Hedges $g$ | $P(\mu)$ | $P(sd)$ | $p(g > 0)$ |
|---|---|---|---|
| Vague (Uniformative) | 0 | 10 | .5 |
| Small | .25 | .5 | .69 |
| Medium | .5 | .5 | .84 |
| Large | .8 | .3 | .99 |

**Figure 3: Posterior medians with 95% HDIs for study level effect sizes, modeling the effect size as a function of control type for the intercept only model and the 4 model variants with different priors.**

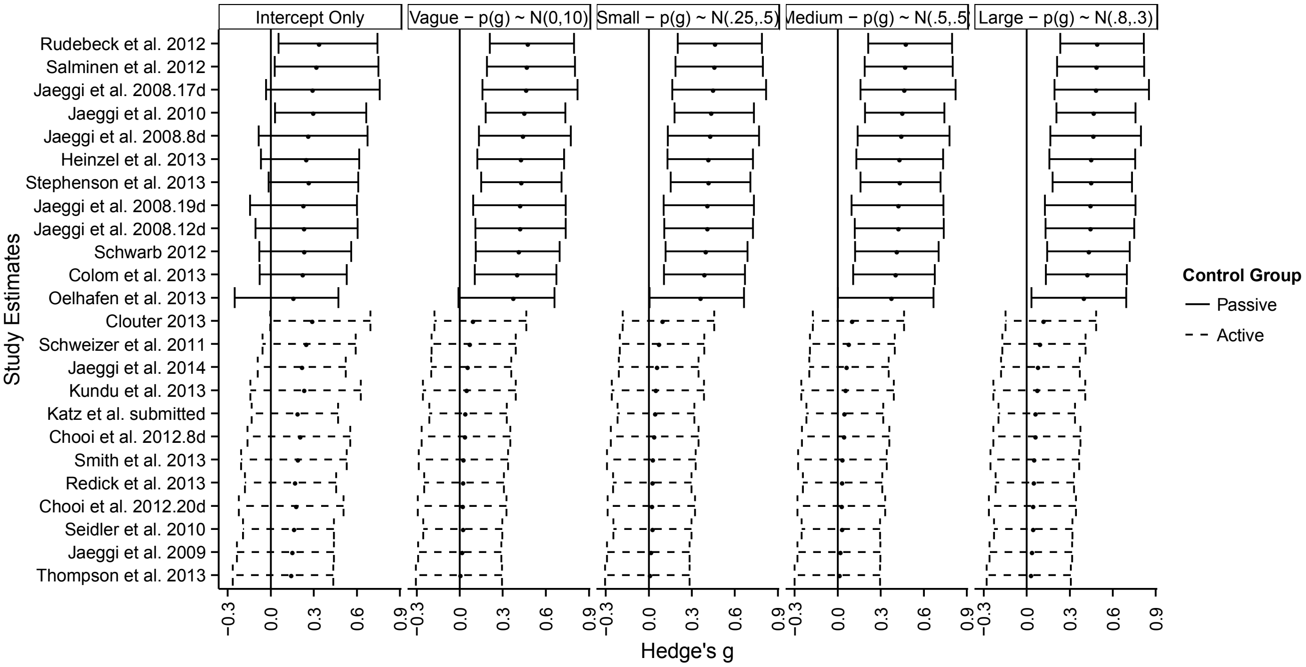


**Figure 4: Posterior medians with 95% HDIs for group level effect sizes, modeling the effect size as a function of control type for the intercept only model and the 4 model variants with different priors.**

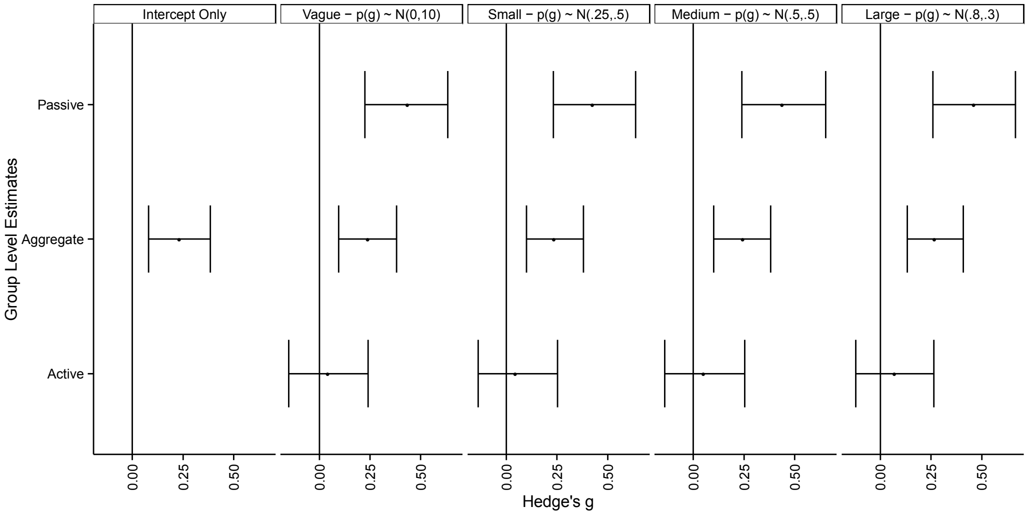


**Figure 5: Posterior medians with 95% HDIs for study level effect sizes, modeling the effect size as an additive function of control type and country of origin for the 4 model variants with different priors.**

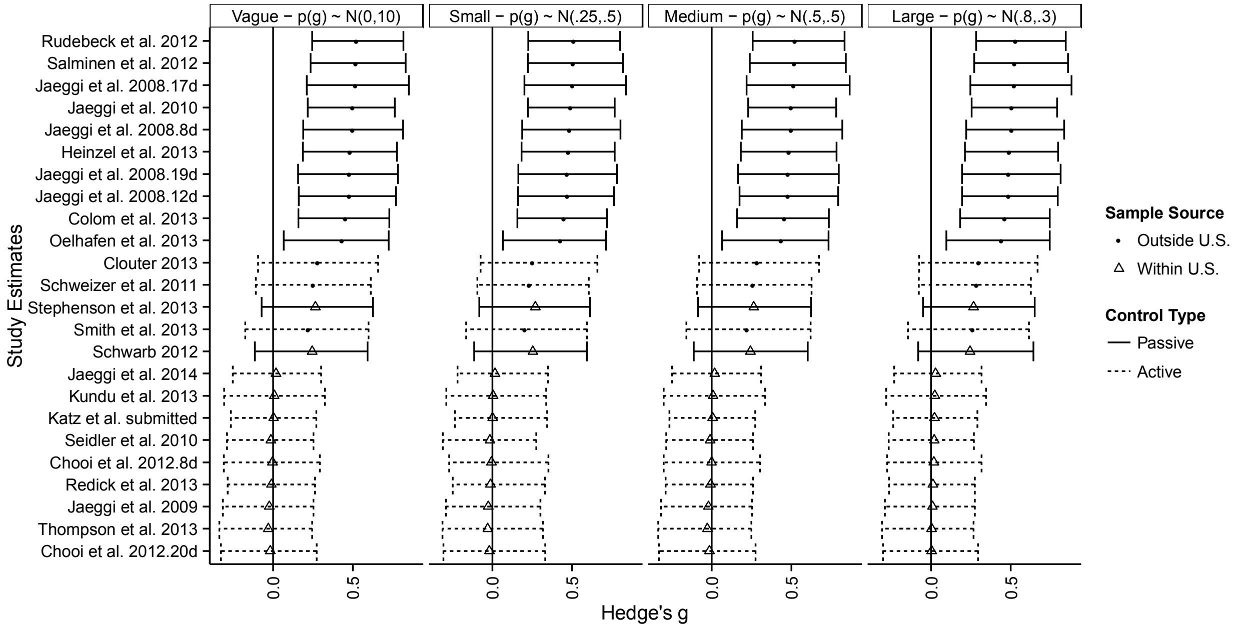
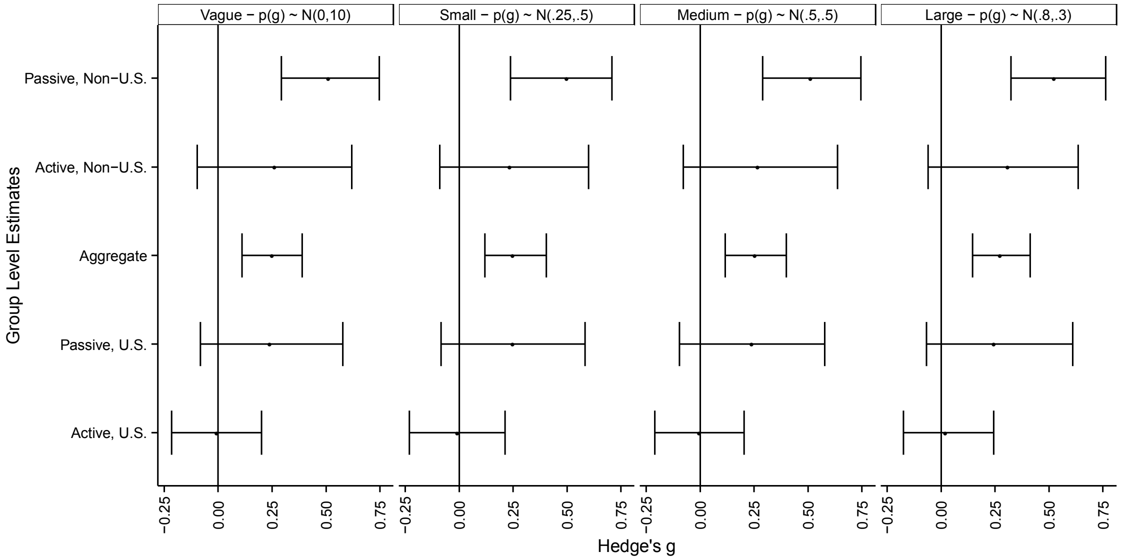


**Figure 6: Posterior medians with 95% HDIs for group level effect sizes, modeling the effect size as a function of control type for the 4 model variants with different priors.**



What is the most appropriate interpretation of these findings? First, it is reasonable to discount the findings of the passive-control studies based on methodological considerations. Because the passive control studies do not control for potential placebo effects, there is no way of discerning whether the effects reflect true training gains or a placebo effect. In fact, the mere size of the BF for the passive control studies should be enough to warrant a critical eye to those studies, especially given the *a priori* uncertainty surrounding the question of whether WM training can improve Gf. This leaves us with the 12 active control studies, for which (a) the Bayes Factors for the individual studies overwhelmingly favor the null, (b) the meta-analytic BF favors the null, (c) the estimated effect sizes are not different from zero, and (d) half of the studies show raw effect sizes indicating a negative effect of transfer.

Second, if one were to interpret the effect sizes of the passive control studies, it would need to be relative to those studies that controlled for placebo effects. Because the passive control studies show a substantial positive effect while the active control studies do not, it seems reasonable to assume that the effects observed in the passive control studies reflects something other than a training effect. The hierarchical Bayesian models suggest a two-factor model for explaining training effects: One factor is the type of experimental design used by the researcher (active versus passive control) and the other is country of origin of the study (USA versus non-USA). We submit that the discrepancy between the active and passive controls is consistent with a placebo effect and we suspect that the effect of country of origin reflects idiosyncratic differences in experimental methods between the USA and non-USA studies. Setting aside specific causal mechanisms for the observed pattern of effect sizes, it is clear that the data reflect two separate data-generating processes, neither of which can be attributed to n-back training.

## Discussion

The results of our re-analysis (and re-interpretation) of the meta analysis of n-back training suggests that to date, the evidence largely fails to support the contention that Gf can be improved through short-term training on n-back. This assertion is supported both at the level of the individual studies and at the aggregate level, and is consistent with several other studies that used different forms of training (e.g., Sprenger et al., 2013; Redick et al., 2013), as well as meta analyses conducted on a broader array of training task-types (Melby-Lervåg & Hulme, 2012). At the same time, however, we note that other meta analyses have recently been completed, some of which seem to support the effectiveness of WM-training (e.g., Karbach & Verhaeghen, 2014, though see Melby-Lervåg & Hulme, 2015), and others that largely fail to do so (Melby-Lervåg, Redick, & Hulme, 2014). Whether a Bayesian analysis of other training tasks would yield similar findings to our analysis of N-back training is an open question, though there is at least one study using Bayesian analyses that showed that transfer from other forms of training to non-trained tasks uniformly favored the null (see Sprenger et al., 2013).

At a more general level, we argue that the evaluation of WM-training effectiveness requires careful attention to detail in the construction of the experimental design, the choice of transfer tasks, and statistical analyses. Although Au et al. (2015) were extremely thorough in collecting studies for inclusion in the meta-analysis, they did not offer a plausible explanation for why the magnitude of the training effect is over 7 times larger (.44 versus .06) when researchers use an experimental design that includes a passive control as opposed to an active control. Choice of experimental design should not moderate the effectiveness of a manipulation, unless of course the design creates the effect through confounding variables.[5] If an effect is contingent on the type of experimental design the researcher uses, then it is the design that is driving the effect, not the experimental manipulation. In the case of the passive control design, participant expectations are confounded with whether they engage in training or not, leaving these studies susceptible to placebo effects masquerading as a training effect.

Along with other recent discussions of placebo effects (Finniss et al., 2010), we suggest that our conclusions should serve as reminder of the possible influence of placebo effects and the need to control for them in intervention studies. An abundance of work now shows that people's expectations can drive everything from pain perception (Atlas & Wager, 2012) and perceptions about migraines (Kam-Hansen et al., 2014), to perceptions regarding the quality of consumer goods (Dougherty & Shanteau, 1999) as well as performance on cognitive tasks (Colagiuri & Boakes, 2010; see also Boot et al., 2013). Given the prevalence of placebo effects, the discrepancy in findings between active and passive experimental designs cannot simply be described as a moderation effect; it has to be fully considered as a possible root cause of the effect. Of course, in the absence of experiments specifically designed to test the placebo effect explanation, it is

[5] It should be noted, however, that choice of experimental design also covaried with whether the study was conducted within the USA or outside the USA. Most of the studies using active controls were conducted within the USA, whereas the majority of the studies conducted outside of the USA used passive controls. While this leaves open the possibility that cultural differences are driving the difference between the active and passive studies, we doubt cultural differences would account for the 7 fold increase in the training effect, especially since the non-USA studies were primarily conducted in Westernized cultures (e.g., Europe).

impossible to definitively state that the difference between active and passive control studies reflects a placebo effect.[6] However, what we can say with some confidence is that if n-back training has a true effect on Gf, then these effects should hold even for studies that use active controls.

From a statistical methodology perspective, the present analysis illustrates the usefulness of the Bayesian approach. First, rather than relying on a p-value to infer the presence or absence of an effect, the Bayesian approach allows one to quantify the strength of the evidence. Individually, studies that find statistically significant effects may not actually provide much evidence for or against the null (see Wetzels et al., 2011). For example, in their analysis of WM training, Chein and Morrison (2010) report a significant effect of complex span training on executive control with a t(38)=1.81, which was reported as significant (p=0.039, one sided). However, assuming a one-sided prior on the effect size, the corresponding BF is only 1.78 (BF = 0.98 for the two-sided test) in favor of the alternative. This is basically uninformative with respect to both the alternative and the null hypothesis. Second, and perhaps more important, the strength of the evidence can be evaluated in relation to any theoretically justified hypothesis, including the null hypothesis. This is important in the domain of WM-training because the main point of disagreement in the literature pertains to whether training leads to

[6] As argued by Hróbjartsson, Kaptchuk, and Miller (2011), there is an appreciable challenge in separating the magnitude of any 'real' placebo effect from variability due to human interaction in an experiment: Due to causal indeterminacy, one cannot simply compare different types of control conditions to infer the presence or absence of placebo effects. In true placebo-control trials, the causal mechanism of the treatment is presumably isolated by virtue of including the placebo control condition. However, the same is not true when comparing a placebo-control with a non-contact control. In these comparisons, there is no way to isolate the effect of the placebo because there are many factors that differ between these conditions (see Hróbjartsson et al., 2011). The problem is even more complicated when comparing placebo controls and no-contact controls drawn from different studies, as it is reasonable to assume that studies that adopt active controls might also adopt other procedures that minimize expectancy or placebo effects.

improvements on non-trained tasks (far transfer), where the null hypothesis is a theoretically meaningful and plausible hypothesis.

It is important to note that the present analysis, as well as that of Au et al. (2015) focuses on transfer to measures of Gf. While we argue that the meta-analysis of n-back training does not support the contention that Gf improves with short-term cognitive training, this does not mean that n-back training does not lead to other forms of transfer: Training on n-back is likely to lead to improvements on other tasks that are similar in design and structure to the n-back task, as demonstrated by Lilienthal, Tamez, Shelton, Myerson, and Hale (2013) and von Bastian & Eschen (in press). However, such transfer effects are neither surprising nor of much practical interest, and neither of these studies found evidence for far transfer. On the other hand, understanding the mechanisms of change on the actual training tasks themselves is an interesting theoretical question (see Harbison, Atkins, & Dougherty, 2015).

In sum, our re-analysis suggests that it is methodological factors, and not the actual n-back training intervention that account for previously observed transfer effects to measures of Gf. Unfortunately, methodological deficiencies in both design and analysis persist in the WM training literature, despite many prior suggestions for remediation (Boot et al., 2013; Tidwell et al., 2014; Shipstead, Redick, & Engle, 2012). One of these areas of methodological deficiencies entails the continued use of passive controls and the other the use of inappropriate analysis techniques that entails correlating training gains with transfer gains (see Tidwell et al., 2014). Further, the continued use of null hypothesis significance testing in this area of research risks overstating the strength of the evidence. While it may very well be the case that other forms of WM training can lead to improvements in general cognitive functions, the meta-analysis presented here and in Au et al. (2015) on n-back training does not provide such evidence.

## Authors note

Michael R. Dougherty, Toby Hamovitz, and Joe W. Tidwell, Department of Psycholology, University of Maryland, College Park, MD 20742. The authors thank Jacky Au and Susan Jaeggi for sharing their data and for providing details of their analysis.

## References

Atlas, L. Y., & Wager, T. D. (2012). How expectations shape pain. *Neuroscience letters*, *520*(2), 140–8. doi:10.1016/j.neulet.2012.03.039

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

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

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), 193–9. doi:10.3758/PBR.17.2.193

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. doi:10.3758/PBR.17.2.193

Chooi, W., & Thompson, L. A. (2012). Working memory training does not improve intelligence in healthy young adults. *Intelligence*, *40*, 531–542.

Clouter, A. (2013). *The Effects of Dual n-back Training on the Components of Working Memory and Fluid Intelligence: An Individual Differences Approach.* (MS), Dalhousie University, Halifax, Nova Scotia.

Colagiuri, B., & Boakes, R. a. (2010). Perceived treatment, feedback, and placebo effects in double-blind RCTs: An experimental analysis. *Psychopharmacology*, *208*(3), 433–41. doi:10.1007/s00213-009-1743-9

Colom, R., Román, F. J., Abad, F. J., Shih, P. C., Privado, J., Froufe, M., … Jaeggi, S. M. (2013). Adaptive n-back training does not improve fluid intelligence at the construct level: Gains on individual tests suggest that training may enhance visuospatial processing. *Intelligence*, *41*(5), 712–727. doi:10.1016/j.intell.2013.09.002

Dougherty, M. R., & Shanteau, J. (1999). Averaging expectancies and perceptual experiences in the assessment of quality. *Acta psychologica*, *101*(1), 49-67.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current directions in psychological science*, *11*(1), 19–23. doi:10.1111/1467-8721.00160

Finniss, D. G., Kaptchuk, T. J., Miller, F., & Benedetti, F. (2010). Biological, clinical, and ethical advances of placebo effects. *Lancet*, *375*(9715), 686–95. doi:10.1016/S0140-6736(09)61706-2

Harbison, J. I., Atkins, S. M., & Dougherty, M. R. (2015). Working memory training improves recollection: A cognitive model-based analysis of WM-training. *Manuscript submitted for publication.*

Harbison, J.I., Atkins, S.M., & Dougherty, M.R. (2015) Working memory training improves recollection: A cognitive model-based analysis of WM-training. Manuscript under revision for resubmission.

Heinzel, S., Schulte, S., Onken, J., Duong, Q., Riemer, T. G., Heinz, A., … Rapp, M. A. (2013). Working memory training improvements and gains in non-trained cognitive tasks in young and older adults. *Aging, Neuropsychology, and Cognition*, *21*(2), 146–73. doi:10.1080/13825585.2013.790338

Hróbjartsson, A., Kaptchuk, T.J., & Miller, F. G. (2011). Placebo effect studies are susceptible to response bias and other types of biases. *Journal of Clinical Epidemiology, 64,* 1223-1229.

Jaeggi, S. M., Buschkuehl, M., & Jonides, J. (2009, June 2-3). Working Memory Training and Transfer. Presented at the Annual ONR Contractor's Meeting, Arlington, VA. As cited in Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: a meta-analysis. *Psychonomic bulletin & review, 22,* 366-377. doi:10.3758/s13423-014-0699-x

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), 6829–33. doi:10.1073/pnas.0801268105

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short- and long-term benefits of cognitive training. Proceedings of the National Academy of Sciences, 108, 10081 – 10086. doi: 10.1073/pnas.1103228108

Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (2014). The role of individual differences in cognitive training and transfer. *Memory & cognition*, *42*(3), 464–80. doi:10.3758/s13421-013-0364-z

Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y., 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

Jeffreys, H. (1961). *Theory of probability*. Oxford University Press.

Kam-Hansen, S., Jakubowski, M., Kelley, J. M., Kirsch, I., Hoaglin, D. C., Kaptchuk, T. J., & Burstein, R. (2014). Altered placebo and drug labeling changes the outcome of episodic migraine attacks. *Science translational medicine*, *6*(218), 218ra5. doi:10.1126/scitranslmed.3006175

Kaptchuk, T. J., Friedlander, E., Kelley, J. M., Sanchez, M. N., Kokkotou, E., Singer, J. P., … Lembo, A. J. (2010). Placebos without deception: A randomized controlled trial in irritable bowel syndrome. *PloS one*, *5*(12), e15591. doi:10.1371/journal.pone.0015591

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

Kass, R. E., Raftery, A. E. (1995). Bayes Factor. *Journal of the American Statistical Association, 90,* 773-795.

Katz, B., Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (under review). Money can't buy you fluid intelligence (but it might not hurt either): The effect of compensation on transfer following a working memory intervention. As cited in Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: a meta-analysis. *Psychonomic bulletin & review, 22,* 366-377. doi:10.3758/s13423-014-0699-x

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. *The journal of neuroscience*, *33*(20), 8705–15. doi:10.1523/JNEUROSCI.5565-12.2013

Lilienthal, L., Tamez, E., Shelton, J. T., Myerson, J., & Hale, S. (2013). Dual n-back training increases the focus of attention. *Psychonomic Bulletin & Review, 20,* 135-141.

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental psychology*, *49*(2), 270–91. doi:10.1037/a0028228

Melby-Lervåg, M., Redick, T., & Hulme, C. (2014). Working Memory Training Does Not Improve Performance on Measures of Intelligence or Other Measures of "Far Transfer": Evidence from a Meta-Analytic Review. Manuscript submitted for publication.

Melby-Lervåg, M., & Hulme, C. (2015). There is no convincing evidence that working memory training is effective: A reply to Au, et al. (2015) and Karbach and Verhaeghen (2014). Manuscript submitted for publication.

Morey, D., Rouder, J. N., & Jamil, T. (2014). BayesFactor: Computation of Bayes factors for common designs. R package version 0.9.8. http://CRAN.R-project.org/package=BayesFactor

Oei, A. C., & Patterson, M. D. (2013). Enhancing cognition with video games: A multiple game training study. *PloS one*, *8*(3), e58546. doi:10.1371/journal.pone.0058546

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), 2781–90. doi:10.1016/j.neuropsychologia.2013.08.012

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Raftery, A. (1995). Bayesian model selection in social research. *Sociological methodology*. Retrieved from https://www.stat.washington.edu/raftery/Research/PDF/socmeth1995.pdf

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), 359–79. doi:10.1037/a0029082

Rode, C., Robson, R., Purviance, A., Geary, D. C., & Mayr, U. (2014). Is working memory training effective? A study in a school setting. *PloS one*, *9*(8), e104796. doi:10.1371/journal.pone.0104796

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic bulletin & review*, *18*(4), 682–9. doi:10.3758/s13423-011-0088-7

Rouder, J. N., Morey, R. D., & Province, J. M. (2013). A Bayes factor meta-analysis of recent extrasensory perception experiments: comment on Storm, Tressoldi, and Di Risio (2010). *Psychological bulletin*, *139*(1), 241–7. doi:10.1037/a0029008

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), 225–37. doi:10.3758/PBR.16.2.225

Rudebeck, S. R., Bor, D., Ormond, A., O'Reilly, J. X., & Lee, A. C. H. (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*(June), 166. doi:10.3389/fnhum.2012.00166

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

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

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. 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

Sprenger, A. M., Atkins, S. M., Bolger, D. J., Harbison, J. I., Novick, J. M., Chrabaszcz, J. S., … Dougherty, M. R. (2013). Training working memory: Limits of transfer. *Intelligence*, *41*(5), 638–663. doi:10.1016/j.intell.2013.07.013

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

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

Tidwell, J. W., Dougherty, M. R., Chrabaszcz, J. R., Thomas, R. P., & Mendoza, J. L. (2014). What counts as evidence for working memory training? Problems with correlated gains and dichotomization. *Psychonomic bulletin & review*, *21*(3), 620–8. doi:10.3758/s13423-013-0560-7

von Bastian, C., C. & Eschen, A. (2015). Does working memory training need to be adaptive? *Psychological Research*. DOI 10.1007/s00426-015-0655-z

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, *14*(5), 779–804.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives On Psychological Science*, *6*(3), 291-298. doi:10.1177/1745691611406923

Willcutt, E. G., Doyle, A. E., Nigg, J. T., Faraone, S. V, & Pennington, B. F. (2005). Validity of the executive function theory of attention-deficit/hyperactivity disorder: a meta-analytic review. *Biological Psychiatry*, *57*(11), 1336–46. doi:10.1016/j.biopsych.2005.02.006

# Bayesian Models

## 1 Data

All data for the models (Tables 1,2) were taken from Au et al 2014, excepting *study.idx*.

| Variable | Description |
|---|---|
| **study.idx** | Unique id corresponding to a particular study |
| **author** | Reference associated with study |
| **n1** | Sample size for experimental condition |
| **n2** | Sample size for control condition |
| **control** | 0 = Passive control; 1 = Active control |
| **USA.location** | 0 = Study conducted outside U.S.; 1 = Study conducted w/in U.S. |
| **Hedges.g** | Effect size |
| **var.g** | Estimated variance of Hedges $g$ |

Table 1: Description of variables used for modeling effect sizes.

| study.idx | author | n1 | n2 | control | USA.location | Hedges.g | var.g |
|---|---|---|---|---|---|---|---|
| 1 | Chooi et al. 2012.8d | 9 | 15 | 1 | 1 | 0.02 | 0.17 |
| 2 | Chooi et al. 2012.20d | 13 | 11 | 1 | 1 | -0.21 | 0.16 |
| 3 | Clouter 2013 | 18 | 18 | 1 | 0 | 0.65 | 0.11 |
| 4 | Colom et al. 2013 | 28 | 28 | 0 | 0 | 0.21 | 0.07 |
| 5 | Heinzel et al. 2013 | 15 | 15 | 0 | 0 | 0.39 | 0.13 |
| 6 | Jaeggi et al. 2010 | 46 | 43 | 0 | 0 | 0.55 | 0.07 |
| 7 | Jaeggi et al. 2009 | 22 | 21 | 1 | 1 | -0.19 | 0.09 |
| 8 | Jaeggi et al. 2008.8d | 8 | 8 | 0 | 0 | 0.64 | 0.24 |
| 9 | Jaeggi et al. 2008.12d | 11 | 11 | 0 | 0 | 0.28 | 0.17 |
| 10 | Jaeggi et al. 2008.17d | 8 | 8 | 0 | 0 | 1.11 | 0.26 |
| 11 | Jaeggi et al. 2008.19d | 7 | 8 | 0 | 0 | 0.22 | 0.24 |
| 12 | Jaeggi et al. 2014 | 51 | 27 | 1 | 1 | 0.18 | 0.08 |
| 13 | Katz et al. submitted | 36 | 27 | 1 | 1 | 0.05 | 0.06 |
| 14 | Kundu et al. 2013 | 13 | 13 | 1 | 1 | 0.34 | 0.31 |
| 15 | Oelhafen et al. 2013 | 14 | 15 | 0 | 0 | -0.28 | 0.13 |
| 16 | Redick et al. 2013 | 24 | 29 | 1 | 1 | -0.05 | 0.07 |
| 17 | Rudebeck et al. 2012 | 27 | 28 | 0 | 0 | 0.76 | 0.08 |
| 18 | Salminen et al. 2012 | 20 | 18 | 0 | 0 | 0.82 | 0.11 |
| 19 | Schwarb 2012 | 22 | 22 | 0 | 1 | 0.26 | 0.09 |
| 20 | Schweizer et al. 2011 | 29 | 16 | 1 | 0 | 0.35 | 0.09 |
| 21 | Seidler et al. 2010 | 29 | 27 | 1 | 1 | -0.07 | 0.07 |
| 22 | Smith et al. 2013 | 10 | 9 | 1 | 0 | -0.16 | 0.19 |
| 23 | Stephenson et al. 2013 | 82 | 26 | 0 | 1 | 0.42 | 0.08 |
| 24 | Thompson et al. 2013 | 20 | 19 | 1 | 1 | -0.28 | 0.10 |

Table 2: Dataset used for modeling of effect sizes.

# 2 Models

## 2.1 Intercept Only

The purpose of this model was to replicate, within a Bayesian framework, the meta-analysis contained in Au et al 2014, and included no predictors.

### 2.1.1 Statistical Model

Here we model the observed effect sizes, $g_i$, as normally distributed with means $\gamma_i$ and observed standard deviations $s_i$. We assume that these effect sizes are drawn from a common distribution of population level effects that are normally distributed with mean $\mu$ and standard deviation $\sigma$, both with vague hyper-parameters.

$$g_i \sim N(\gamma_i, s_i) \tag{1}$$

$$\gamma_i \sim N(\mu, \sigma) \tag{2}$$

$$\mu \sim N(0, 100) \tag{3}$$

$$\sigma \sim U(0, 100) \tag{4}$$

### 2.1.2 Stan Code

All parameter estimation was conducted using the **rstan** version of the Stan modeling language (Stan Development Team, 2014) for the R programming language (R Core Team, 2014). In this, and subsequent sections, we provide the Stan models and R code used to run each model.

```
model.intercept <- '
data {
  int<lower=0> N;
  real g[N];
  real<lower=0> s[N];
  int<lower=0> nu;
}
parameters {
  real gamma[N];
  real mu;
  real<lower=0, upper=100> sigma;
}
model {
  mu ~ normal(0, 100);
  gamma ~ normal(mu, sigma);
  g ~ normal(gamma, s);
}
'
intercept.data <- list(g  = mydata$Hedges.g,
                       s  = sqrt(mydata$var.g),
                       N  = length(mydata$Hedges.g))
samples.intercept <- stan(model_code=model.intercept,
                          data=intercept.data,
                          iter=1e5,
                          chains=3,
                          thin=1)
```

## 2.2 Control Type

The purpose of these four models was to determine the effect of control type {passive, active} on the aggregate effect size.

### 2.2.1 Statistical Model

$$g_i \sim N(\gamma_i, s_i) \tag{5}$$

$$\gamma_i \sim N(X\theta_j, \sigma),\ j = 1, 2 \tag{6}$$

$$\theta_j \sim N(\mu_j, \tau_j) \tag{7}$$

$$\mu \sim N(0, 100) \tag{8}$$

$$\sigma \sim U(0, 100) \tag{9}$$

$$\tag{10}$$

This model extends the intercept only model by regressing $\gamma_i$ on the model matrix $X$ to estimate parameters $\theta_j$. In this case $X$ includes a column of 1's for an intercept and an effect coded column for control type, yielding an intercept $\theta_1$ that models the aggregate effect size across predictors, $\theta_1$, and the relative difference of the control groups from the aggregate effect size $\theta_2$. Each $\theta_j$ is distributed normal with mean $\mu_j$ and standard deviation $\tau_j$.

We estimated four instances of this general model. For each instance, we assumed the mildly informed prior $\theta_2 \sim N(\mu_2 = 0, \tau_2 = 5)$. This represents the assumption that there is no strong prior expectation that the control groups should differ from the aggregate effect size. The prior on $\theta_1$ was varied across all four instances to represent the prior beliefs of: 1) No expected effect, $N(\mu_1 = 0, \tau_1 = 10)$; 2) A small expected effect, $N(\mu_1 = .25, \tau_1 = .5)$; 3) A medium expected effect, $N(\mu_1 = .5, \tau_1 = .5)$; and 4) A large expected effect, $N(\mu_1 = .8, \tau_1 = .3)$.

### 2.2.2 Stan Code

```
model.control <- '
data {
  int<lower=0> N;
  real g[N];
  real<lower=0> s[N];
  int<lower=1> K;
  matrix[N,K] X;
  real pr_mu;
  real pr_sd;
}
parameters {
  real gamma[N];
  vector[K] theta;
  real<lower=0, upper=100> sigma;
}
model {
  theta[1] ~ normal(pr_mu,pr_sd);
  theta[2] ~ normal(0,pr_sd);
  gamma ~ normal(X*theta, sigma);
  g ~ normal(gamma,s);
}
'
```

```
pars <- data.frame(mu=c(0,.25,.5,.8), sd=c(10,.5,.5,.3))
models.control <- foreach(i = 1:nrow(pars)) %do% {
  control.data <- list(g  = mydata$Hedges.g,
                       s  = sqrt(mydata$var.g),
                       N  = length(mydata$Hedges.g),
                       X  = X,
                       K  = 2,
                       pr_mu = pars[i,"mu"],
                       pr_sd = pars[i,"sd"])
    stan(model_code=model.control,
       data=control.data,
       iter=1e5,
       chains=3,
       thin=1,
       refresh=-1)
}
```

## 2.3 Control Type + Country of Origin

The purpose of these four models was to determine the additive effect of control type {passive, active} and country of study {Non-U.S., U.S.} on the aggregate effect size.

### 2.3.1 Statistical Model

$$g_i \sim N(\gamma_i, s_i) \tag{11}$$

$$\gamma_i \sim N(X\theta_j, \sigma),\ j = 1 \dots 3 \tag{12}$$

$$\theta_j \sim N(\mu_j, \tau_j) \tag{13}$$

$$\mu \sim N(0, 100) \tag{14}$$

$$\sigma \sim U(0, 100) \tag{15}$$

$$\tag{16}$$

This model extends the control model by including an additional dichotomous covariate in $X$ that identifies whether a study was conducted within or outside of the United States.

We estimated four instances of this general model. For each instance, we assumed the mildly informed independent priors $\theta_2, \theta_3 \sim N(\mu_2 = 0, \tau_2 = 5)$. This represents the assumption that there is no strong prior expectation that either the control groups or country of study should differ from the aggregate effect size. The prior on $\theta_1$ was varied across all four instances to represent the prior beliefs of: 1) No expected effect, $N(\mu_1 = 0, \tau_1 = 10)$; 2) A small expected effect, $N(\mu_1 = .25, \tau_1 = .5)$; 3) A medium expected effect, $N(\mu_1 = .5, \tau_1 = .5)$; and 4) A large expected effect, $N(\mu_1 = .8, \tau_1 = .3)$.

### 2.3.2 Stan Code

```
model.country <- '
data {
  int<lower=0> N;
  real g[N];
  real<lower=0> s[N];
  int<lower=1> K;
  matrix[N,K] X;
  int<lower=0> nu;
  real pr_mu;
  real pr_sd;
}
parameters {
  real gamma[N];
  vector[K] theta;
  real<lower=0, upper=100> sigma;
}
model {
  theta[1] ~ normal(pr_mu,pr_sd);
  theta[2] ~ normal(0,5);
  theta[3] ~ normal(0,5);
  gamma ~ normal(X*theta, sigma);
  g ~ normal(gamma,s);
}
'

models.country <- foreach(i = 1:nrow(pars)) %do% {
  country.data <- list(g  = mydata$Hedges.g,
                       s  = sqrt(mydata$var.g),
                       N  = length(mydata$Hedges.g),
                       X  = X,
                       K  = 3,
                       pr_mu = pars[i,"mu"],
                       pr_sd = pars[i,"sd"])
    stan(model_code=model.country,
       data=country.data,
       iter=5e5,
       chains=3,
       thin=1,
       refresh=-1)
}
```