Abstract

Bediou et al. (2018a) meta-analyzed the effects of action-game training interventions, concluding, after a trim-and-fill adjustment for publication bias, that such interventions produce modest transfer of training to performance on a broad range of cognitive tasks (*g* = 0.34; after bias adjustment, *g* = 0.23). We conducted a meta-analysis on the same set of studies, and after applying stronger and more appropriate adjustments for publication bias, we found little evidence for transfer of training; instead, the results seem largely attributable to publication bias. Moreover, an erratum on the original paper, triggered by an earlier draft of this manuscript, acknowledged that different outcomes from the same participants appear in multiple publications. This revelation supports the possibility of publication bias: Some outcomes might not have been reported at all. The erratum models the overlaps between publications, but this underreporting is still a problem for claims of transfer of training. More compelling, independent, and transparent evidence is needed before concluding that action video game training transfers to performance on other cognitive tasks.

In a meta-analysis of randomized controlled trials that compared an action video game intervention to a control group, Bediou et al. (2018a; hereafter BAMTGB) report evidence for modest benefits of game training for a broad range of cognitive outcome measures.[[1]](#footnote-2) Such broad transfer is notable given limited evidence for transfer of training from other cognitive training interventions (e.g., Melby-Lervåg, Redick, & Hulme, 2016; Sala & Gobet, 2017) as well as strong historical evidence that training tends to provide little benefit for tasks other than those specifically trained (see Simons et al., 2016).

The BAMTGB analysis concludes in favor of transfer of training, but that conclusion is premature. Here we reexamine the set of studies included in their meta-analysis, and after more thoroughly accounting for publication bias, we find the evidence for action-game training effects is weak, with small or possibly null effects on cognitive function (see also a recent meta-analysis by Sala, Tatlidil, & Gobet, 2018 drawing a similar conclusion).

Furthermore, an erratum to the BAMTGB analysis (Bediou et al., 2018b) revealed that outcome measures from participants in some intervention studies appeared in separate publications and were incorrectly treated as independent in the meta-analysis.[[2]](#footnote-3) At the study level, this “salami slicing” approach to publishing does not adequately correct for multiple testing; the first paper published from an intervention does not correct for tests of outcome measures published later, and vice versa. Moreover, without a complete accounting of all measures tested, it is likely that only a subset of the measured outcomes have yet been published.

**Publication Bias**

Publication bias is a threat to the accuracy of meta-analytic results. If statistically significant effects are more likely to be reported, published, and retrieved for meta-analysis than non-significant results, meta-analysis will overestimate the true effect. In the presence of publication bias, even a null effect can appear robust. Appropriate application and interpretation of adjustments for publication bias are necessary to avoid drawing overly firm conclusions about the size or presence of a true effect.

BAMTGB report potential publication bias in the analyzed studies, cautioning that the published literature finds larger effects than would be expected in a meta-analysis of all conducted studies. They report an asymmetrical funnel plot and a significant Egger test (*b* = 4.96, *z* = 3.95, *p* < .001), suggesting publication bias or small-study effects.[[3]](#footnote-4) BAMTGB reported PET and PEESE adjustments for this bias (Stanley, 2017; Stanley & Doucouliagos, 2014) and acknowledged that the estimates indicated bias. However, they chose not to focus on these estimates because they “sometimes result in estimated effects in the opposite direction [which may] suggest limitations with the PET and PEESE approaches to publication bias detection and correction” (Bediou et al., 2018a, p 95). Instead, they relied on trim-and-fill (Duval & Tweedie, 2000) to adjust for publication bias. Because this bias adjustment did not eliminate the observed overall effect, BAMTGB concluded in favor of training benefits in their significance statement.

The PET and PEESE results must be considered more thoroughly, not because the negative estimates are correct, but because they are wrong in a way that indicates severe publication bias. The large, negative PET-adjusted effect size of *g* = -1.69 may be a sign that the true effect has been badly overestimated. In simulations, the combination of publication bias and questionable research practices (e.g., underreporting of outcomes, optional stopping) tends to cause PET to dramatically undershoot the true effect, mistaking null effects for significantly negative ones (Carter, Schonbrodt, Gervais, & Hilgard, 2017). Even though PEESE is usually biased upwards when the null is true, it too returned a negative estimate (*g* = -0.53), again suggesting severe publication bias.

The use of a trim-and-fill adjustment is likely inadequate. In the presence of bias, trim-and-fill slightly improves estimates relative to unadjusted meta-analysis, but it does not adequately adjust for publication bias to recover a true null effect, and Type I error rates remain inflated (Carter et al., 2017; Moreno et al., 2009; see also [Simonsohn, Simmons, & Nelson](http://datacolada.org/30), 2014). Given its tendency to mistake a true null effect for a significant positive effect in the presence of publication bias, trim and fill should not be used to conclude in favor of an effect when there are other signs of substantial publication bias.

However, there are limitations to PET and PEESE adjustments for publication bias. PET and PEESE assume no relationship between effect size and standard error, attributing any such relationship to bias. Such a relationship can exist for benign reasons rather than bias, however. For example, some approaches to calculating the standard error of *g* are a function of *g* itself, with higher values of *g* creating greater standard errors. This assumption can cause PET and PEESE to overestimate the degree of bias and to adjust too much. A small-study effect can also be caused by confounds, rather than bias, say if the larger studies used cheap, ineffective manipulations and smaller studies used expensive, potent ones.

To avoid the specific weaknesses of PET and PEESE, we examined whether other publication bias adjustments, namely *p*-uniform (van Assen, van Aert, & Wicherts, 2015) and selection modeling (Vevea & Hedges, 1995), yield estimates more in line with those of PET-PEESE or trim-and-fill. Unfortunately, these bias adjustments are not designed, in their standard implementations, to handle studies with multiple outcome measures. We adapted these adjustments by using bootstrapping, randomly selecting one outcome from each study, and then applying random-effects meta-analysis, *p*-uniform, and a selection model. This process was repeated 1,000 times to explore the variability in the estimates with different sets of outcomes when randomly selecting one outcome from each study. 95% CIs are defined as the 2.5th and 97.5th percentile of bootstrapped estimates.[[4]](#footnote-5) Although the unadjusted random-effects estimate was medium in size and statistically significant (*g* = 0.50 [0.33, 0.64]), *p*-uniform and selection modeling estimated a small and nonsignificant effect (*p*-uniform, *g* = 0.11, [-0.81, 0.70]; selection modeling, *g* = 0.15 [-0.10, 0.43]). Like PET and PEESE, these adjustments suggest that the uncorrected effect size estimate is inflated due to publication bias and suggest the possibility that the true effect is small or null.

We note the limitation that the CIs are quite wide, especially using *p*-uniform, indicating that some benefits cannot be ruled out. Still, given the limited support for transfer of cognitive training, (Melby-Lervåg et al., 2016), we feel the burden of proof rests on those claiming efficacy of game training.

Table 1 summarizes the results of several publication bias adjustments across three different ways of accommodating the clustering of effects within studies. (See the supplement for results using other, less appropriate bias correction methods.) Note that PET and PEESE also yield substantially reduced and nonsignificant estimates of game training effects across various strategies for handling dependency within clusters.

**BAMTGB’s lab effect and publication bias.**

The overall effect estimate reported by BAMTGB was qualified by a large lab effect: Research by Bavelier and colleagues (several of whom are co-authors of BAMTGB) yielded substantially larger effects (*g* = 0.92 [0.76, 1.08], *p* < .001) than studies by other groups (*g* = 0.22 [-0.01, 0.45], *p* = .054). In the original article, BAMTGB suggest that the disparity results from the longer training durations used by Bavelier and colleagues. However, their moderation analysis provides only weak evidence of an effect of training duration (pp. 91-92), and in the corrected analysis in the erratum, the moderation is no longer statistically significant (*b* = 0.016 [-.004, .036], *p* = .089). Training durations do not, therefore, explain the pronounced laboratory effect.

Although a lab effect might result from differences in research design, it also could follow from differences in publication practices. We assessed whether differences in publication bias might contribute to the lab effect by adding the PET adjustment for small-study effects to the laboratory moderator analysis. Adding this adjustment for publication bias dramatically reduced the laboratory effect from *b* = 0.72, *t*(4.28) = 7.35, *p* = .001, *ω*2 = 0.096, *τ*2 = 0.027 without adjustment to *b* = 0.20, *t*(5.36) = 1.85, *p* = .119, *ω*2 = 0.060, *τ*2 = 0 with adjustment. The PET adjustment also further reduced the apparent effect of training duration from *b* = 0.016, *t*(3.47) = 2.35, *p* = .089, *ω*2 = 0.108, *τ*2 = 0.061 without adjustment to *b* = 0.007, *t*(4.32) = 1.66, *p* = .168, *ω*2 = 0.059, *τ*2 = 0 with adjustment. Adding the PET adjustment also accounted for all of the observed between-cluster variability (*τ*2). That is, after adjusting for publication bias, there is no observed heterogeneity across clusters, meaning that there is little evidence for effects of lab or of training duration. Taken together, these analyses suggest that the larger effects in studies reported by Bavelier and colleagues than by other research groups might result from greater publication bias, not differences in research design (e.g., longer training durations).

The observed lab effect could also be caused by a bias against training effects in non-Bavelier lab studies. This seems unlikely given that small-study effects suggestive of bias are detected in the non-Bavelier studies as well. Using robust variance estimation, the Egger test among non-Bavelier studies is significant, *b* = 5.79, *t*(4.3) = 5.04, *p* = .006, and the PET-adjusted estimate is negative, *g* = -1.89, *t*(4.54) = -3.98, *p* = .013. Thus, while the lab effect suggests greater overestimation among Bavelier-group studies, studies by other groups also show evidence of publication bias.

**Unreported overlap between studies**

In the initial version of their meta-analysis, BAMTGB treated partially independent groups of participants as if they were fully independent. That is, some effect sizes that were dependent were instead treated as independent. This risks overstating the precision of the effect size and giving undue weight to trials with participants published in multiple, partially-independent articles.

In response to an earlier version of our manuscript, BAMTGB issued an erratum to their article. In that erratum, they reported a reanalysis of their data that appropriately treated partially-independent samples as belonging to the same cluster. This correction is a more appropriate and conservative statistical approach.

However, readers are still left to wonder about the extent of overlap in the published literature. An inspection of the corrected dataset reveals a surprising number of overlapping samples from the Bavelier group.[[5]](#footnote-6) We say surprising because none of the articles that reported distinct outcome measures from an intervention documented that other outcomes from the same participants were reported elsewhere. Similarly, none of the published papers made clear the full set of outcome measures tested.

An earlier draft of the erratum, posted publicly on the Open Science Framework (Bediou, 2018), identified cases of multiple publication from the same dataset. The draft identifies three reported experiments drawn from one sample, saying “data from the same subjects (different tasks) were reported in more than one study and should thus have been assigned the same cluster.”

Another section of the draft erratum, labeled “Subject Overlap”, indicated broad and unreported overlap across multiple papers:

Effect sizes from the below intervention studies from the Bavelier lab were treated as cases of partial overlap because participants were run during successive but distinct summer waves of training and each summer the trained groups were administered overlapping but not identical tasks […] note that the exact degree of overlap between subjects included in some training studies of the Bavelier lab is impossible to determine accurately as these studies were run at the University of Rochester before 2009; This lab was closed when Bavelier moved to the University of Geneva and unfortunately the records available in Geneva do not include the level of detail necessary to ascertain exact percent overlap).

Neither disclosure appears in the published erratum, which says only “In the original publication, cases of partial overlap were treated as independent. A more conservative approach is to code these effects as dependent, which is done here.”

This failure to disclose overlap harms the scientific process by overstating the amount of independent evidence acquired. As the Office of Research Integrity says, “dividing a study into smaller segments must always be done with full transparency, showing exactly how the data being reported in the later publication are related to the earlier publication. […] Salami slicing can lead to a distortion of the literature by leading unsuspecting readers to believe that data presented in each salami slice (i.e., journal article) are independently derived from a different data collection effort or subject sample.” (Office of Research Integrity, n.d.). We note that the analyses in the original publication treated the data in this manner, implying that the outcome measures were from independent studies when they were not.

This form of overlap can also contribute to publication bias. Readers and reviewers cannot evaluate the results of an intervention until the results for all outcome measures have been published. If some outcome measures are not yet published, as seems likely for non-significant ones, then the effect sizes for those that were published are likely inflated due to the statistical significance filter on publication. For conclusions relying on null-hypothesis significance testing, omission of outcome measures can substantially inflate false positive rates (Simmons et al., 2011). Without complete reporting of all outcomes from an intervention, *p*-values become uninterpretable because it is impossible to control for multiple testing to maintain a fixed familywise error rate.

In addition to the statistical harms, one may also be concerned about the fairness of this practice. An article presented as a new trial might be evaluated more favorably than an analysis of outcomes collected in a previously-reported study. Failure to disclose overlap among samples in studies submitted for publication may thereby provide an unfair advantage in peer review. We are also concerned by the prospect of “overlapping but not identical tasks.” If the treatment and control groups performed different sets of tasks, then the experiments are confounded. If this is the case, then failure to disclose such a weakness would also mislead peer reviewers.

**Accounting for overlapping studies.** We detail below the overlaps revealed by the updated dataset, the earlier draft of the erratum, and our close read of summary statistics in the published articles. Specifically, we note cases in which multiple publication of identical or overlapping participants went unreported.

Three articles separately reporting effects of action game training on visual acuity (Green & Bavelier, 2007), multiple object tracking (Green & Bavelier, 2006a, Cognition), and UFOV (Green & Bavelier, 2006b, JEP:HPP) actually reported different outcomes from the same subjects. None of these articles mentioned that results came from the same intervention.

Similarly, three publications reporting action game training effects on backward masking (Li, Polat, Scalzo, and Bavelier, 2010), contrast sensitivity (Li, Polat, Makous, and Bavelier, 2009), and visual motion discrimination (Green, Pouget, and Bavelier, 2010) were based on overlapping subsets of the same sample of participants. Again, these overlaps are not described in the original articles. (Green et al., 2010 mentions “subjects underwent 50 hours of training as well as several experiments unrelated to the ones at hand,” but it is never indicated what those experiments were or that those experiments were reported in other publications.) Even more perplexing, the subjects in the two training studies reported in a single article (Li et al. 2009, studies 2 and 4) apparently overlapped, with the earlier erratum draft flagging study 4 as “Partial overlap with all” of the other papers reporting results from the same waves of data collection.

The draft erratum also notes overlap in game training effects on the perception of a Gabor signal from noise (Bejjanki et al. 2014) and task-switching (Green et al., 2012). The 2014 publication makes no mention of overlap with the 2012 publication. The 2012 paper does state: “Subjects completed two experimental blocks [of task switching]...as well as several other tasks unrelated to the current paper (e.g., motion discrimination, visual search, contrast detection – however note: the data presented here was acquired over the course of 3 separate training studies – and thus the unrelated tasks are not identical in all subjects)” (Green et al. 2012, pg. 992). However, it is not clear which three separate training studies are included or where those outcomes are reported.

Although not documented in either version of the erratum, the other outcomes mentioned by Green et al (2012) might have appeared in earlier papers. Under the same training conditions (30 hrs of *Unreal* / *Call of Duty* or *Tetris* training), Li et al. (2009) reported effects on contrast sensitivity and Green et al. (2010) reported effects on motion discrimination.[[6]](#footnote-7) If this speculation is correct, then the samples analyzed in Li et al. (2009), Li et al. (2010), Green et al. (2010), and Green et al. (2012) were at least partially overlapping. (The reported sample sizes differ across articles, so the samples in each paper may overlap but are not identical.)

An accurate estimate of the benefits of action video games requires a more complete accounting of the collected samples and outcomes. Even the initial draft of the erratum that reported some of the overlap across papers did not provide a full accounting of all the conducted trials and collected outcomes. Future studies should fully report all outcome measures and explicitly describe any overlap with previous studies. For already published research, we hope that authors will expand on the earlier draft of the erratum and provide a more complete accounting of the overlap among results reported across separate papers: How many fully-independent RCTs have been conducted and what outcomes were collected in each? If samples reported in multiple papers overlapped partially or completely, which participants contributed to each outcome?

**Underreporting of outcomes may cause publication bias and the observed lab effect**

Consistent with the possibility that differences in publication bias account for the lab effect reported by BAMTGB, articles from Bavelier and colleagues report an average of 1.6 outcomes (SD=0.8) per experiment whereas papers from other laboratories average 4 outcomes (SD=2.5; see Figure 1). Collecting only one or two outcome measures in an intervention seems unlikely, and even 4 seems low. Training interventions are expensive and time consuming, so most such studies include batteries of outcome measures. For example, the ACTIVE trial of cognitive training in older adults collected 10 proximal outcomes, 6 primary outcomes, and 5 secondary outcomes (see Jobe et al., 2001). The relative dearth of reported outcomes in video game intervention studies suggests that many of the papers included in the meta-analysis might not have fully reported all of the outcome measures that were collected as part of the reported intervention. If papers are more likely to report significant than non-significant outcomes, underreporting in one set of studies would result in inflated effect sizes for those studies. This would be consistent with the observed lab effect as well as with publication bias in both subsets of studies.

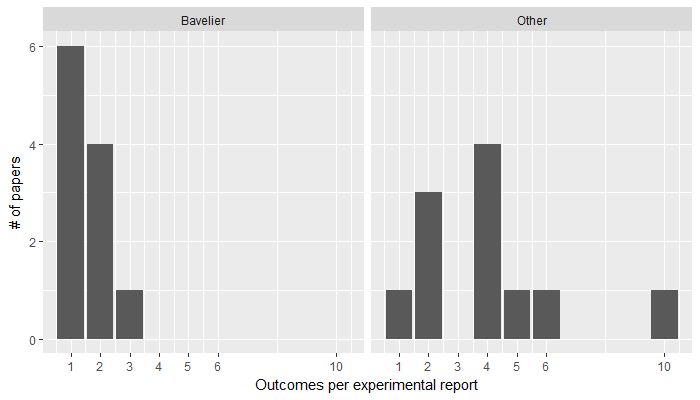
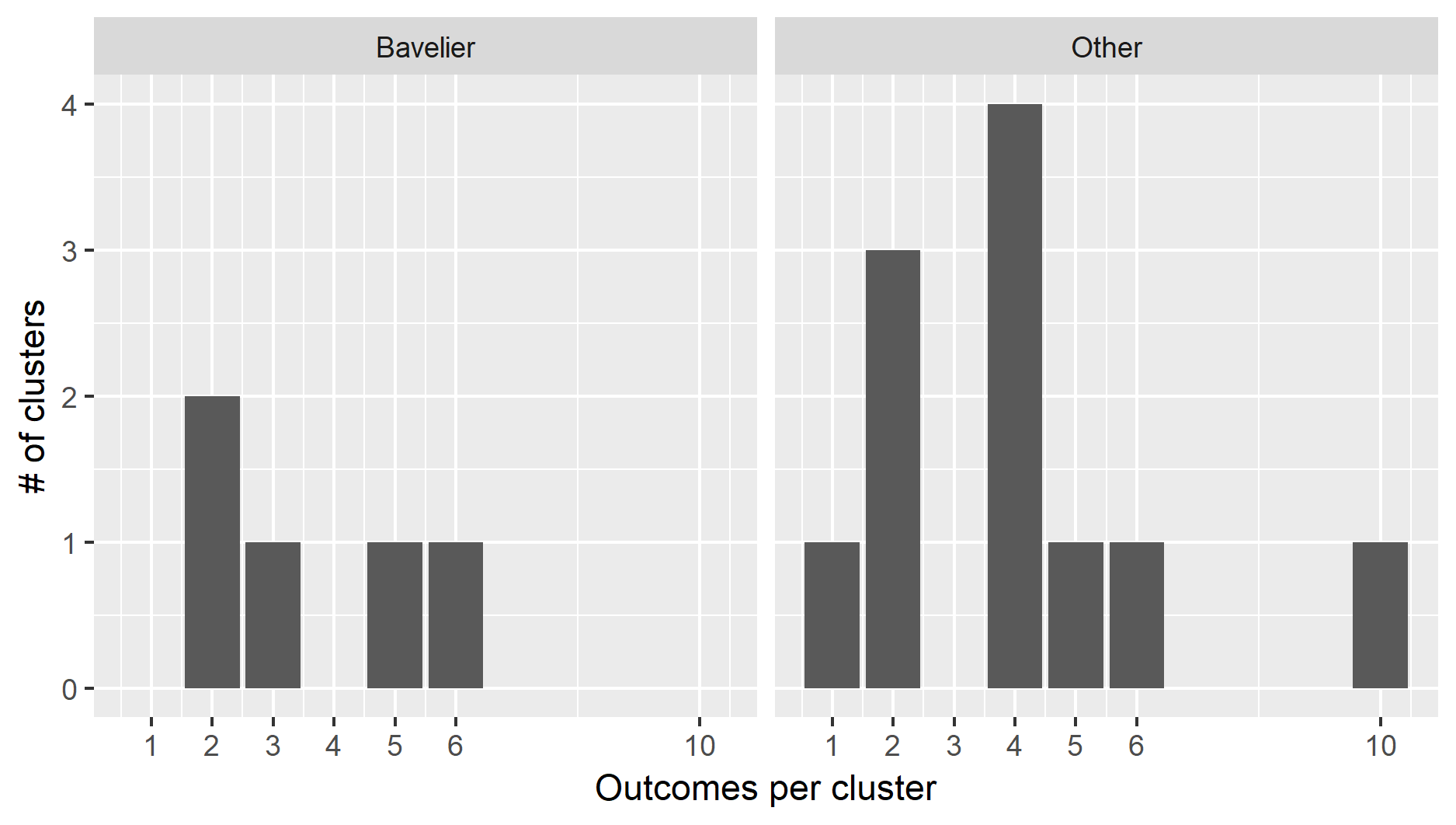


Figure 1. Number of papers (Y axis) in each subset of studies reporting a given number of outcomes (X axis). The subset of papers in BAMTGB by Bavelier and colleagues typically reported fewer outcomes than those by other laboratories.

The erratum reanalyzed the data with fewer clusters, grouping together papers that reported separate outcomes from the same intervention. If we reconstruct Figure 1 using the number of outcomes collected per cluster rather than per reported experiment, the count from Bavelier and company is more similar to that of other labs (*M* = 3.6, *SD* = 1.82) (Figure 2). This pattern is consistent with many outcomes being collected with each intervention and only selected outcomes being published.



**Figure 2.** Number of clusters of (partially) dependent samples reporting a given number of outcomes. After clustering together outcomes from overlapping samples that were reported in separate papers, the count of collected outcomes per study is less markedly different.

**Summary**

BAMTGB acknowledged the possibility of publication bias, but after the trim-and-fill adjustment, they concluded in favor of significant benefits from action game training. However, trim-and-fill is inadequate as an adjustment for the substantial publication bias that meta-regression techniques suggest is present (Carter et al., 2017; Stanley, 2017). Following other more suitable adjustments for publication bias, the meta-analytic results are consistent with a small effect or no effect of action game training on other outcome measures. PET and PEESE estimates were negative, consistent with a small or null effect and strong bias, and bootstrapped *p*-uniform and selection modeling yielded small, non-significant results.

The strong lab effect is consistent with differences in publication bias and is not explained by differences in training duration. The earlier draft of the published erratum documents overlaps between subjects across papers that had not been reported previously, but this was not reported in the published erratum. This overlap, combined with the small number of outcomes reported in each paper and the lack of acknowledgment of the overlap in the published literature, suggests selective publication of results from each intervention. It is unclear whether other significant outcomes are yet to be published or whether non-significant outcomes have gone unreported.

In sum, BAMTGB’s conclusion in favor of training benefits appears premature in the face of substantial publication bias, uncertainty about the number of distinct, independent interventions among the meta-analyzed results, and the potential censoring of outcomes that did not yield significant results. Although game training might transfer to other cognitive tasks, the studies synthesized by BAMTGB do not support that conclusion; the meta-analyzed data are also consistent with a combination of substantial publication bias and no effect of action games on cognition. Greater transparency in the reporting of interventions and of outcomes within interventions is needed.

**References**

Bediou, B. (2018) Meta-analysis of the impact of action video games on cognition. Retrieved May 22, 2018; <https://osf.io/w8xcd/download?version=1&displayName=MA_Intervention_erratum-2018-05-22T17%3A54%3A11.403680%2B00%3A00.pdf>

Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Green, C. S., & Bavelier, D. (2018a). Meta-analysis of action video game impact on perceptual, attentional, and cognitive

skills. *Psychological Bulletin, 144*, 77-110. [doi:10.1037/bul0000130](http://psycnet.apa.org/doi/10.1037/bul0000130)

Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Green, C. S., & Bavelier, D. (2018b). Correction: Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. *Psychological Bulletin,* *144*, 978-979. doi: 10.1037/bul0000168

Carter, E., Schönbrodt, F., Gervais, W. M., & Hilgard, J. (2017). Correcting for bias in psychology: A comparison of meta-analytic methods. Preprint available at https://psyarxiv.com/9h3nu/

Cheung, S. F., & Chan, D. K. (2014). Meta-analyzing dependent correlations: An SPSS macro and an R script. *Behavioral Research Methods, 46*, 331-345. doi:10.3758/s13428-013-0386-2

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel plot based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 276-284. doi:10.1111/j.0006‑341X.2000.00455.x

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature, 423*, 534-537. doi:10.1038/nature01647

Green, C. S., & Bavelier, D. (2006a). Enumeration versus multiple object tracking: The case of action video game players. *Cognition, 101*, 217-245. doi:10.1016/j.cognition.2005.10.004

Green, C. S., & Bavelier, D. (2006b). Effect of action video games on the spatial distribution of visuospatial attention. *Journal of Experimental Psychology: Human Perception and Performance, 32*, 1465-1478. doi:10.1037/0096-1523.32.6.1465

Green, C. S., & Bavelier, D. (2007). Action-video-game experience alters the spatial resolution of vision. *Psychological Science, 18*, 88-94. doi:10.1111/j.1467‑9280.2007.01853.x

Green, C.S., Pouget, A., & Bavelier, D. (2010). Improved probabilistic inference as a general learning mechanism with action video games. *Current Biology, 20*, 1573-1579. [doi:10.1016/j.cub.2010.07.040](https://doi.org/10.1016/j.cub.2010.07.040)

Green, C. S., Sugarman, M. A., Medford, K., Klobusicky, E., & Bavelier, D. (2012). The effect of action video game experience on task-switching. *Computers in Human Behavior, 28*, 984-994. doi:10.1016/j.chb.2011.12.020

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*, 39-65. doi:10.1002/jrsm.5

Jobe, J. B., Smith, D. M., Ball, K., Tennstedt, S. L., Marsiske, M., Willis, S. L., ... & Kleinman, K. (2001). ACTIVE: A cognitive intervention trial to promote independence in older adults. *Controlled clinical trials, 22*, 453-479. doi:10.1016/S0197-2456(01)00139-8

Li, R., Polat, U., Makous, W., & Bavelier, D. (2009). Enhancing the contrast sensitivity function through action video game training. *Nature Neuroscience, 12*, 549-551. doi:10.1038/nn.2296

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of far-transfer: Evidence from a meta-analytic review. *Perspective on Psychological Science, 11*, 512-534. doi:10.1177/1745691616635612

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC medical research methodology, 9*, 2. doi:10.1186/1471-2288-9-2

Office of Research Integrity. (no date). *Redundancy, publication overlap, and other forms of duplication.* Retrieved Jan 1, 2019 from https://ori.hhs.gov/plagiarism-15

Sala, G., & Gobet, F. (2017). Does far transfer exist? Negative evidence from chess, music, and working memory training. *Current Direction in Psychological Science, 26*, 515-520. doi:10.1177/0963721417712760

Sala, G., Tatlidil, K. S., & Gobet, F. (2018). Video game training does not enhance cognitive ability: A comprehensive meta-analytic investigation. *Psychological Bulletin, 144*, 111-139. doi:10.1037/bul0000139

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

Simons, D. J., Boot, W. R., Charness, N., Gathercole, S.E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. L. (2016). Do “brain-training” programs work? *Psychological Science in the Public Interest, 17*, 103-186. doi:10.1177/1529100616661983

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). p-Curve and effect size: correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9*, 666-681. doi:10.1177/1745691614553988

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359-1366. doi.org/10.1177/0956797611417632

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science, 8*, 581-591. doi:10.1177/1948550617693062

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5*, 60-78. doi:10.1002/jrsm.1095

van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015) Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, 20(3), 293-309. doi: 10.1037/met0000025

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, 60*, 419-435. doi:10.1007/BF02294384

**Acknowledgments**

Special thanks to John Sakaluk for consultation regarding treatment of dependent effect size estimates. Thanks also to Dan Ispas, Olivia Cody, and Christopher Engelhardt for reading drafts of the commentary. JH, DS, WB, and GS conceptualized, wrote, and edited this commentary. JH and GS performed the statistical analyses. WB and DS documented the possible overlap across clusters reported in the meta-analysis. GS is a JSPS International Research Fellow.

**Supplement**

The studies analyzed by BAMTGB involve multiple outcomes nested within studies. Although a number of methods exist to account for such dependency of outcomes, not all of them are compatible with various techniques used to account for publication bias. This supplement discusses these combinations of estimation methods and publication bias adjustments, noting which are most appropriate and robust.

**Robust Variance Estimation.** For their unadjusted meta-analytic estimate, BAMTGB use Robust Variance Estimation (RVE; Hedges et al., 2010) which places an upper limit on the weight each study can receive. By adding the standard error or the variance of the effect size as a moderator, it is possible to adjust for publication bias using PET and PEESE. However, RVE is not compatible with trim-and-fill, *p*-uniform, or selection modeling approaches to adjusting for publication bias.

**Averaging within studies.** Another way to model the dependency between effect sizes within a study is to average across all effect sizes within a cluster. In essence, averaging treats all outcomes from an intervention as if they were perfectly correlated with each other. When this assumption is incorrect, as it almost always is, sampling error may be overestimated and heterogeneity may be underestimated (Cheung & Chan, 2014; Schmidt & Hunter, 2015). Still, it is not an uncommon approach, and it is this approach BAMTGB use in their publication bias analyses.

The averaging approach is compatible with trim-and-fill, PET, and PEESE. It is not recommended for use with *p*-uniform and selection modeling; although it can be done, the assumptions of *p*-uniform and selection modeling are likely violated when averaging. Those approaches assume outcomes are published according to their individual *p*-values rather than according to the *p*-value corresponding to the average of several outcomes. Although this supplement provides these estimates, we do not recommend interpreting them.

**Cheung and Chan’s correction for merged effect sizes.** Like the averaging analysis used by BAMTGB, the Cheung and Chan (2014) procedure merges the dependent effect sizes into one effect size (the average of the dependent effect sizes), but it accounts for this dependence by using an adjusted sample size that that falls between the original sample size (*N*) and *N×k* (where *k* is the number of the effect sizes; see samplewise-adjusted-individual estimates in Cheung & Chan, 2014). Those adjusted Ns are then used to calculate the variance for the merged effect sizes, addressing the potential shortcomings in the estimation of sampling error and heterogeneity when using an average effect size.

As with averaging, this approach is compatible with PET and PEESE meta-regression. However, it is not recommended for *p*-uniform or selection modeling, as the Cheung-and-Chan-corrected *p*-values do not match the *p*-values used in publication decisions.

**Treating outcomes as independent.** Although it is possible to treat each outcome measure as entirely independent from all other outcome measures, doing so is not recommended as it assumes a correlation of zero between outcomes. It is possible to apply all bias-adjustment methods when using this (flawed) aggregation strategy.

**Bootstrapping.** Bootstrapping involves randomly selecting one outcome from each study, conducting a meta-analysis, and then repeating that process many times. Doing so allows an assessment of the sensitivity of the meta-analytic estimate to the choice of one outcome from each cluster. Bootstrapping meets the assumptions of *p*-uniform and selection models because only one effect size is considered per study, and that one effect size retains its original *p*-value. Note, though, that bootstrapping treats all outcomes reported for a study as theoretically equivalent tests of the primary hypothesis. If some effects are more important than others for theoretical reasons, it would be better to select those for analysis. We apply *p*-uniform and selection modeling through bootstrapping.

**Supplementary results.** Table 1 in the main text provides the results for recommended combinations of dependency modeling and adjustment for bias. For completeness, we provide the other calculable results in Table S1, even though they are not recommended.

Regardless of how dependency is modeled, PET and PEESE still indicate overestimation and a lack of significant evidence for an effect. Estimates from *p*-uniform and three-parameter selection modeling vary depending on how dependency is modeled. Estimates range from *g* = 0.11 to *g* = 1.37. Again, we do not think these adjustments are suitable for use with these approaches to handling dependency, as the assumptions of the model are violated, and we encourage interpretation of the results in the main text.

1. In this commentary, we focus exclusively on the experimental intervention results because those are the ones that permit causal inferences about the effects of video game training on cognitive performance. All of the core claims about benefits of game training depend on those studies. Some of the issues we raise might apply to the cross-sectional results as well, but we have not re-analyzed those results. [↑](#footnote-ref-2)
2. The erratum recomputed the meta-analysis after clustering such non-independent findings together. All of our calculations use the corrected BAMTGB dataset from their erratum. Data and code are available at https://osf.io/dhejx/?view\_only=b318140cb10a4a1c9b13e83b781d9f24 [↑](#footnote-ref-3)
3. This test is based on averaging outcomes within studies. It is also significant when using Robust Variance Estimation, *b* = 5.28, *t*(6.84) = 7.40, *p* < .001. [↑](#footnote-ref-4)
4. Because neither selection models nor *p*-uniform were designed to be used with multiple outcomes, we consulted two experts in meta-analysis. One expert told us that bootstrapping was a decent idea for selection modeling, while the other told us that bootstrapping was a bad idea for *p*-uniform. The reader is cautioned that bias-adjustment through bootstrapping these methods is a novel approach. We provide these estimates so as to explore the robustness of the results. [↑](#footnote-ref-5)
5. The erratum for BAMTGB acknowledges the overlaps by reassigning overlapping samples to the same cluster. However, it does not make these overlaps explicit. An earlier draft of that erratum, posted publicly on the Open Science Framework (retrieved May 22, 2018; https://osf.io/w8xcd/download?version=1&displayName=MA\_Intervention\_erratum-2018-05-22T17%3A54%3A11.403680%2B00%3A00.pdf), included a section labeled “Subject Overlap” that was not included in the published erratum. We quote from that section in the main text. [↑](#footnote-ref-6)
6. Note that the 2012 paper did not specify that the other outcomes were reported in the earlier papers, and neither of the earlier papers mentioned specific outcomes reported elsewhere. Whether the overlap in participants across these papers is partial or total is unclear in the published record. [↑](#footnote-ref-7)