UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Hilgard, J., et al. (2019). Overestimation of Action-Game Training Effects: Publication Bias and Salami Slicing. *Collabra: Psychology,* 5(1): 30. DOI: https://doi.org/10.1525/collabra.231

ORIGINAL RESEARCH REPORT

Overestimation of Action-Game Training Effects: Publication Bias and Salami Slicing

Joseph Hilgard[*], Giovanni Sala[†], Walter R. Boot[‡] and Daniel J. Simons[§]

Does playing action video games improve performance on tests of cognitive ability? A recent meta-analysis (Bediou et al., 2018a) summarized the available evidence and concluded that it can. Their analysis, however, did not adequately correct for publication bias. We re-analyzed the same set of studies with more appropriate adjustments for publication bias and found minimal evidence for transfer of training to cognitive ability measures. Instead, it is possible that there are little or no benefits, just publication bias — the exclusion of non-significant results from the published literature. That bias may be the cause of a lab effect reported in the original meta-analysis. The meta-analysis showed that studies from the Bavelier lab (the senior author of the meta-analysis) reported larger effects than other labs. We show that many of these original studies distributed different outcomes from the same or highly overlapping sets of participants across publications without noting the overlap. This salami-slicing might contribute to the extent of publication bias in the literature. More compelling, independent, and transparent evidence is needed before concluding that action video game training transfers to performance on other cognitive tasks.

Keywords: video game training; action game training; cognitive training; meta-analysis; publication bias; salami slicing

Can playing action video games improve your ability to perform other basic cognitive tasks? In a seminal study (Green & Bavelier, 2003), participants who played a first-person shooter game for 10 hours showed bigger improvements on measures of spatial attention than did to those who played a control game. Such broad transfer was 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).

This excitement generated by this finding inspired many subsequent tests of the effectiveness of videogame training on cognition, and it contributed to growing interest in potential benefits of cognitive training more broadly (Simons et al., 2016). A recent meta-analysis of randomized controlled trials on the benefits of action game training (Bediou et al., 2018a; hereafter BAMTGB), conducted by a team including the authors of that seminal study (Green and Bavelier), concluded that action games have modest benefits for a broad range of cognitive outcome measures.

We believe that conclusion is premature. We re-analyzed their meta-analysis with the same set of studies to more fully evaluate the influence of publication bias (i.e., exclusion of non-significant findings from the literature). Although BAMTGB observed indications of publication bias, they dismissed those measures and argued that they did not threaten their conclusions. When using more appropriate adjustments, we find evidence of problematic levels of publication bias. In our re-analysis, we show that the existing evidence is weak and that it realistically could reflect a small or null benefit of game training coupled with publication bias. Whether or not games benefit cognition remains unclear because of publication bias and underreporting.

The concerns about publication bias are amplified by two additional factors, addressed below: 1) The meta-analysis reported a large lab effect, with studies from Bavelier and colleagues (co-authors on the meta-analysis) showing much larger effects than those from other laboratories. 2) An erratum to the Bediou et al. article revealed a possible source of this lab effect (Bediou et al., 2018b): different outcome measures from the same participants in some Bavelier lab intervention studies appeared in separate publications.[1] In a later section of this paper, we document this previously unreported "salami slicing" and explain how it can lead to overestimated benefits of action-game training.

[*] Illinois State University, Normal, Illinois, US

[†] Osaka University, Osaka, JP

[‡] Florida State University, Tallahassee, Florida, US

[§] University of Illinois Urbana-Champaign, US

Corresponding author: Joseph Hilgard (jhilgard@gmail.com)

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.

A number of statistical procedures are available to test and adjust for publication bias. (See the Glossary in **Table 1**. For more details about these methods, see the papers cited in **Table 1** or the summaries in Carter, Schönbrodt, Gervais, & Hilgard, 2019.) No adjustment for publication bias is perfect, but adjusting for publication bias is preferable to relying on an unadjusted random-effects meta-analysis. An unadjusted meta-analysis assumes that all relevant data have been included; this assumption is often violated. Simulation studies show that adjustments can help reduce error in the presence of publication bias (Carter et al., 2019; Moreno et al., 2009; Simonsohn, Nelson, & Simmons, 2014; van Assen, van Aert, & Wicherts, 2015). For example, an unadjusted meta-analysis of ego-depletion experiments estimated a sizeable effect ($d = 0.62$), but a re-analysis applying the precision-effect test and precision-effect estimate with standard errors (PET-PEESE) adjustment estimated the mean effect to be approximately zero (Carter & McCullough, 2014). The PET-PEESE prediction was consistent with the results of a subsequent multi-site Registered Replication Report (Hagger et al., 2016).

Because each adjustment has weaknesses, a better approach involves a sensitivity analysis that considers results across multiple adjustment methods (Carter et al., 2019; McShane, Böckenholt, & Hansen, 2016) and evaluates the degree to which estimates are either robust or sensitive to variations in the method of adjustment. However, different adjustments are flawed in different ways. Under the same conditions, some suffer from upward bias and others suffer from downward bias. Consequently, the adjustments will not always converge on the same estimate.

BAMTGB applied three approaches to estimating the consequences of publication bias: the Egger test, trim and fill, and PET-PEESE. They observed a significant Egger test, indicating the existence of a small-study effect consistent with publication bias. The PET and PEESE adjustments for this small study bias both were so severe that they returned significant *negative* estimates. That is, when extrapolating from the included studies to a prediction of what the effect would be for a study with an infinite sample size, both estimated an effect indicating significantly greater improvements for the control group than for game training. Because this conclusion was implausible, the authors chose to dismiss these estimates as "suggest[ing] limitations with the PET and PEESE approaches to publication bias detection and correction" (Bediou et al., 2018a, p 95). They instead relied exclusively on the medium effect size estimate ($g = 0.48$) from trim and fill.

Unfortunately, this trim-and-fill estimate is likely to be too large because trim and fill generally does not adjust enough when there is publication bias (Carter et al., 2019; Simonsohn, Simmons, & Nelson, 2014; van Assen et al., 2015). Moreover, PET can return large and significant negative estimates when a set of studies includes both publication bias and questionable research practices (e.g., underreporting of outcomes, optional stopping; Carter et al., 2019). Although the PET estimate of $g = -1.69$ is unlikely to represent the true effect, it is consistent with a badly overestimated true effect from the random-effects meta-analysis and the trim-and-fill adjustment. PEESE is usually biased *upwards* when the null is true, but it too returned a negative estimate ($g = -0.53$), again suggesting severe publication bias. These negative estimates indicate the presence of strong publication bias, meaning that the trim-and-fill adjustment is likely insufficient.

Another potential indication of bias is the large laboratory effect reported by BAMTGB. Studies from the laboratory group of Bavelier (a senior author on BAMTGB) found significantly larger effects than studies from other laboratories. In the original version of the meta-analysis, the authors argued that the larger effects they had observed could be attributed to their use of longer training durations. In the published erratum, though, moderation by training duration was not significant, $p = .089$. An alternative explanation, one we discuss below, is that published results from the Bavelier laboratory were more likely to overestimate the true effect size.

In the next section, we provide a more thorough analysis of publication bias in the BAMTGB meta-analysis. In the subsequent section, we examine how differences in publication practices might contribute to publication bias and explain the laboratory effect reported by BAMTGB.

Examining Publication bias in Bediou et al. (2018a)

Method

We first reproduced the estimates from Bediou et al. (2018a) and then applied all the adjustments for publication bias described in **Table 1**. We also probed the robustness of the moderation by laboratory (Bavelier and non-Bavelier) and examined whether that moderation could be explained by differences in training duration or differences in publication bias. Data and code are available at https://osf.io/dhejx/.

Dependency between outcomes within studies. One challenge of adjusting for publication bias in this meta-analysis is that studies have multiple outcomes. Simple meta-analytic models assume a single effect size per study; a more sophisticated model is required to handle multiple effect sizes per study. BAMTGB did so using robust variance estimation (Hedges, Tipton, & Johnson, 2010), a method for modeling and estimating the dependence between outcomes within studies. BAMTGB applied the PET and PEESE adjustments in concert with this robust variance estimation. However, the trim-and-fill adjustment is incompatible with robust variance estimation, so the authors computed their trim-and-fill

Table 1: Glossary of bias-adjustment techniques.

Adjustment method	Summary	Further information
Egger test	Tests for small-study effects by regressing observed effect sizes against their standard errors. Since standard error (sample size) does not cause effect size, no relationship is expected in the absence of publication bias. A negative slope indicates bigger effect sizes for smaller studies. This can be caused by publication bias: small studies only reach statistical significance when they have overestimated the true effect size, whereas large studies can be published without such overestimation. In the absence of compelling reasons to expect bigger effects for smaller studies, a significant slope suggests evidence of publication bias.	Egger, Davey Smith, Schneider, & Minder (1997)
Trim and Fill	Adjusts for small-study effects (small studies showing bigger effects) by imputing studies with the "missing" negative or small effects. Stronger small-study effects result in more imputed studies and a stronger reduction in the effect size estimate. Although popular, this adjustment often does not adjust strongly enough when there is publication bias.	Duval & Tweedie (2000)
PET	Like the Egger Test, PET relies on regression of effect size against a measure of study precision. It considers the intercept rather than the slope. This estimates the effect size that would be predicted from a linear extrapolation to a perfectly precise study (infinite sample size). This linear model assumes that all studies face equal publication bias regardless of their sample size. PET tends to underestimate the size of non-null effects.	Stanley & Doucouliagos (2013)
PEESE	PEESE adopts the same approach as PET, except that the extrapolation uses a quadratic, rather than a linear, relationship with study precision. This quadratic model assumes that publication bias is stronger among small studies, which must overestimate the effect to get published, and weaker among large studies, which are well-powered enough to avoid the file drawer. PEESE tends to overestimate the size of null effects.	Stanley & Doucouliagos (2013)
PET-PEESE	PET and PEESE are often used in combination, as a conditional PET-PEESE estimator. Because PET is biased downwards when the null is false, and PEESE is biased upwards when the null is true, PET-PEESE attempts to apply the estimator that is more likely to be accurate. It first applies the PET adjustment; if the estimate is significant, it switches to PEESE. This tends to inherit PET's downward bias, since PET has poor power to reject the null.	Stanley & Doucouliagos (2013)
p-uniform	*p*-uniform estimates the true effect size using the distribution of *p* values for only those studies that produced a statistically significant result. When the null is true, the distribution of statistically significant *p* values is expected to be uniform. When there is a true positive effect, the distribution of *p* values should be right skewed, with more low *p* values than high *p* values. The extent of the right skew is proportional to the power of the average statistical power of the studies, and the approach provides an estimate of the true effect that would yield that level of skew. It is fundamentally similar to *p*-curve.	van Assen, van Aert, & Wicherts (2015)
Three-parameter selection modeling (3PSM)	This approach models publication bias with a parameter representing how much less likely a nonsignificant result is to be published than a significant result. The other two parameters represent the estimated bias-adjusted mean effect and the estimated heterogeneity of the effects.	Hedges & Vevea (1996)

estimate by first averaging across outcomes within a study and then applying the adjustment. We used the same approach in our PET, PEESE, and trim-and-fill adjustments to reproduce their analyses.

The *p*-uniform and 3PSM adjustments (**Table 1**) are incompatible with robust variance estimation and with averaging outcomes within studies. Both methods assume that studies are either published or censored based on a single primary outcome's *p* value rather than on the average *p* value across outcomes. These models are expected to perform poorly when using the average of *p* values (Simonsohn, Nelson, et al., 2014), so the standard recommendation is to choose the single most appropriate outcome from each study.

Unfortunately, we cannot know which outcome is the most appropriate, nor can we know which outcome or outcomes were subjected to selection bias. We considered creating a specification curve (Simonsohn, Simmons, & Nelson, 2015), running *p*-uniform and selection modeling on every possible choice of outcome per study. This approach was infeasible, though, as it would require more than 1.5 billion analyses. Instead, we chose to use bootstrapping, which randomly samples from this space of 1.5 billion possible analyses. In this way, bootstrapping gives all outcomes equal weight, which is an appropriate compromise when a single primary outcome cannot be identified. We bootstrapped 1,000 samples to explore the variability in the estimates caused by the choice of outcome from each study. 95% CIs are defined as the 2.5th and 97.5th percentile of bootstrapped estimates.

Results

Table 2 summarizes the results of our analyses with and without adjustment for publication bias using the updated data set provided in the erratum. (See the supplement for a sensitivity analysis using other, less appropriate combinations of bias adjustment and study-level aggregation methods.) We reproduced the unadjusted random-effects, PET, and PEESE estimates of the main effect of game training calculated in the authors' R Markdown document provided in the online data repository (MA_Intervention_erratum.Rmd). Their PET and PEESE calculations mistakenly used the "weights" rather than "modelweights" argument, and we corrected that error in our code. (The differences in the estimates were minor. Other small differences may result from our testing of single moderators rather than multiple moderators.) The unadjusted random-effects estimate was significant and positive (g = 0.34 [0.07, 0.61]), but both PET and PEESE yielded significant and negative estimates for the erratum dataset (PET: g = –1.69 [–2.44, –0.94]; PEESE: g = –0.53 [–0.95, –0.11]) just as they did in the original analysis (Bediou et al., 2018a, Figure S7).

BAMTGB chose to interpret these negative PET and PEESE results as indicating problematic and eccentric behavior of the PET and PEESE estimators. However, our *p*-uniform and 3PSM adjustments were also consistent with a substantial effect of publication bias, with both estimating small and nonsignificant effects (*p*-uniform, g = 0.11, [–0.78, 0.69]; selection modeling, g = 0.16 [–0.09, 0.42]).

None of these estimates, on its own, is necessarily accurate or precise. Most have wide confidence intervals (especially *p*-uniform). Furthermore, many involve a tradeoff between bias and variance, with some minimizing variance but retaining bias (trim and fill) and others reducing bias in favor of noisier estimates. Nevertheless, these estimates collectively indicate a serious and problematic degree of publication bias and the possibility of no effect of game training on cognition. Although we cannot say for certain that there is no benefit, given limited support for transfer of cognitive training more generally (Melby-Lervåg et al., 2016), we feel the burden of proof rests on those claiming efficacy of game training to show that such effects are not an artifact of publication bias.

Lab effect. The overall effect estimate reported by BAMTGB was qualified by a large lab effect: Research from Bavelier laboratory 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). BAMTGB reported a moderation analysis and concluded that the lab effect

Table 2: Bias-adjusted effect size estimates.

Aggregation	Estimator		All labs	Bavelier lab	Other labs
Robust	RE		0.34 [0.07, 0.61]	0.92 [0.76, 1.08]	0.22 [–0.01, 0.45]
Robust	PET	*	–1.69 [–2.44, –0.94]	–0.24 [–1.92, 1.44]	–1.89 [–3.16, –0.63]
Robust	PEESE	*	–0.53 [–0.95, –0.11]	0.38 [–0.23, 0.99]	–0.75 [–1.4, –0.09]
Averaged	RE		0.46 [0.24, 0.68]	0.95 [0.53, 1.37]	0.29 [0.06, 0.51]
Averaged	Trim&Fill		0.28 [0.03, 0.53]	0.85 [0.49, 1.21]	0.17 [–0.05, 0.38]
Bootstrapped	RE		0.49 [0.32, 0.65]	0.91 [0.67, 1.16]	0.33 [0.12, 0.52]
Bootstrapped	SelectionModel	*	0.16 [–0.09, 0.42]	0.63 [0.18, 0.99]	0.12 [–0.12, 0.45]
Bootstrapped	P-uniform	*	0.11 [–0.78, 0.69]	–0.15 [–2.03, 0.89]	0.23 [–0.55, 0.89]

Note: Random-effects estimates not adjusted for publication bias included for comparison. Trim-and-fill estimate included as a reproduction of the authors' analyses. PET and PEESE indicate substantial overestimation of the true effect size, and *p*-uniform and a three-parameter selection model concur.

* We recommend these analyses for interpretation for their demonstrated bias-adjustment ability (Carter et al., 2019). Other entries in this table are presented to reproduce the original BAMTGB estimates or to show the analysis' sensitivity to how multiple outcomes are analyzed.

likely resulted from the longer training durations used in their own studies. The weak moderation effect in the original report was no longer statistically significant in the updated data set provided in the erratum ($b = 0.016$ [–.004, .036], $p = .089$). Differences in training durations do not, therefore, explain the pronounced laboratory effect.

Although a lab effect might result from differences in methods and research design, it also could result from differences in publication practices. We tested this possibility by including the PET bias adjustment in a multiple regression predicting effect size from the lab producing the effect. If the lab effect is caused by differences in research design, then including the PET bias adjustment in the regression should either have little effect on the estimated lab effect or it should strengthen it by removing noise. If the lab effect is caused in part by differences in publication bias, however, then including the PET adjustment for publication bias in the regression will reduce the size of the estimated lab effect by removing explained variance.

Adding the PET adjustment dramatically reduced the laboratory effect from $b = 0.72$, $t(4.28) = 7.35$, $p = .001$, $\omega^2 = 0.096$, $\tau^2 = 0.027$ without adjustment to $b = 0.20$, $t(5.36) = 1.85$, $p = .119$, $\omega^2 = 0.060$, $\tau^2 = 0$ with adjustment. Much of the laboratory effect might be attributable to differences in publication bias. Adding the PET adjustment also accounted for all of the observed variability between study-level effect sizes (i.e., the between-cluster variability, τ^2), meaning that after adjusting for small study biases, the variation in the observed effect sizes between studies can be explained by sampling error alone. Taken together, these analyses suggest that the larger effects in studies reported by Bavelier laboratory might result from greater publication bias, not differences in training duration or other design differences.

One might argue that the Bavelier lab results are accurate and that it is the non-Bavelier labs that have *under*estimated the true effect through selection bias in *favor* of the null. This seems unlikely given that small-study effects suggestive of bias against the null are detected in the non-Bavelier-lab studies as well. The Egger test among non-Bavelier-lab 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-lab studies, studies by other laboratories also show evidence of publication bias and overestimation. This widespread publication bias indicates serious weakness in the evidence for game training benefits.

Unreported overlap between studies

Might differences in reporting practices underlie the apparent difference in publication bias between the Bavelier lab studies and other studies? In this section, we examine how one practice—distributing outcomes from the same study across multiple articles—might account for much of the reported lab effect and lead to an inflated overall estimate of the benefits of game training.

Contemporary standards for transparency, such as those from the Office of Research Integrity, require clear documentation whenever data from the same study are reported in multiple publications:

> [D]ividing 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.)

Salami slicing can contribute to overestimated effects in meta-analysis. If a meta-analysis treats the outcomes reported in each paper as if they were independent of those reported in other papers, then salami slicing of outcomes across papers will lead to a single intervention counting multiple times. That is, the analysis will treat these distributed outcomes as if they were independent intervention studies, implying that there was more evidence than actually exists. And, If those papers happened to report some of the larger effect sizes, the overall effect size estimate will be further inflated. Salami slicing is particularly damaging when individual studies collect many outcomes and only significant outcomes are published, a practice that constitutes a form of *p*-hacking (Simmons, Nelson, & Simonsohn, 2011). Censoring of non-significant outcomes biases the meta-analytic estimate upwards, and when combined with salami slicing, each significant outcome is weighted even more heavily into the overall meta-analytic estimate, amplifying the effects of publication bias.

Independent of the statistical hazards, undisclosed salami slicing might increase the likelihood that a result will be published at all by misleading editors and reviewers. A manuscript presented as a new trial might be evaluated more favorably than one reporting outcomes collected in a previously-reported study.

Salami slicing in action video game training

To what extent might salami slicing contribute to publication bias and the reported lab effect in studies of video game training? Prior to the publication of the BAMTGB meta-analysis, two of us had noted the possibility that data from one trial might have been published in separate articles without clear documentation. We called for more transparency in the extent of overlap (Boot, Blakely, & Simons, 2011). For example, several sets of articles reported experiments with highly similar training methods (same games, same durations, same game improvements), with each reporting only a small number of outcomes.

The original BAMTGB analysis clustered outcome measures from the same study together to account for their non-independence, but they did not cluster outcomes across papers. Although BAMTGB did note that some unspecified papers were not entirely independent, their analysis treated outcomes from different papers as if they were always from different interventions with different participants. In response to an earlier version of this manuscript that we sent to the BAMTGB authors, they issued an erratum that re-analyzed the data by assigning dependent groups of participants across papers to the same cluster. This revised clustering identifies overlapping samples, providing an opportunity to determine which original studies featured undisclosed overlapping samples.

Methods

We compared the clusters in the corrected BAMTGB dataset against the original BAMTGB dataset. In addition, we carefully read the published articles and compared their methods and results, looking for similarities. This analysis was aided by an early draft of the BAMTGB erratum, posted publicly on the Open Science Framework, which we quote from below.[2] This early draft included a section labeled "Subject Overlap" which identified overlaps and documented the extent of those overlaps (complete in some cases, partial in others). This "subject overlap" section was removed in the published version of the erratum, which instead only stated: "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."

Results

The corrected dataset clustered together a number of papers that had been treated as independent in the original meta-analysis, all of which were among the Bavelier lab subset of studies. To our knowledge, this re-clustering is the first time this overlap has been addressed: none of the primary articles reported the full set of outcomes collected, nor did these articles report that other outcome measures from the same participants were previously reported in other published work.

The earlier draft of the erratum text identified three papers that reported outcomes from the same 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" (Bediou et al., 2018b, p. 1). These articles separately report 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). None of these articles mentioned that the reported outcomes and results came from the same sample.

The "Subject Overlap" section also noted other broad, 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). (Bediou et al., 2018b, p. 1)

Based on the erratum draft and our own readings of the papers, we document here several additional sets of overlapping papers. The first set involves three papers reporting four studies of action game training effects on backward masking (Li, Polat, Scalzo, & Bavelier, 2010, Study 3), contrast sensitivity (Li, Polat, Makous, & Bavelier, 2009, Studies 2 and 4), and visual motion discrimination (Green, Pouget, & Bavelier, 2010, Study 3). The erratum indicates that the three Li et al. studies are drawn from subsamples of the Green et al. (2010) experiment. These overlaps are not described in the original articles. (Green et al., 2010, supplement p. 15 mentions "subjects underwent 50 hours of training as well as several experiments unrelated to the ones at hand," but does not indicate what those experiments were or whether "experiments" referred to outcome measures from the intervention that were reported in other publications.)

Another overlapping set, noted in the initial erratum draft, includes outcome measures of task-switching (Green, Sugarman, Medford, Klobusicky, & Bavelier, 2012) and seeing the orientation of a line under visual noise (Bejjanki et al., 2014). The 2014 publication makes no mention of overlap with the 2012 publication. The 2012 paper notes: "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 unclear which three separate training studies are included or where those outcomes are reported. One possibility is that those outcomes had been reported in earlier papers that also reported the results of 30 hours of training with either *Unreal Tournament/Call of Duty* or *Sims 2*. Such effects of training on contrast sensitivity were reported in Li et al. (2009, Study 4), and effects on motion discrimination were reported in Green et al. (2010, Study 3). If these are the outcomes alluded to in Green et al. (2012), this partial overlap was not documented in the first draft of the erratum, was not accounted for by re-clustering in the updated dataset, and is not mentioned in the original papers. (The sample sizes differ across these articles, so the extent of possible overlap is unknown.)

Underreporting of outcomes may explain both bias and the lab effect

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).

How many outcomes are reported in typical action game training articles? Those from Bavelier and colleagues report an average of 1.6 outcomes (*SD* = 0.8) whereas papers from other laboratories average 4 outcomes per experiment (*SD* = 2.5; see **Figure 1**). Collecting only one or two outcome measures in an intervention seems unlikely, and even 4 seems low.

Why so few outcomes? One possibility is that that more outcomes were collected, but those outcomes that did not show a statistically significant difference in improvement between the game training and control conditions were not reported. This selection bias among outcomes within studies would cause overestimation of the effect size. Since the Bavelier lab group reports a greater average improvement per outcome, but fewer outcomes per paper, it is possible that the lab effect could be explained by differences in reporting practices.

Another possibility is that relatively few outcomes are reported per published study because some outcomes are salami-sliced for later publications. If we reconstruct **Figure 1** using the number of outcomes collected per cluster in the updated dataset rather than the number reported for each paper, the count from Bavelier and colleagues is closer to that reported in papers by other labs (*M* = 3.6, *SD* = 1.82) (**Figure 2**).

In the absence of trial registration or disclosure statements about the number of outcomes (e.g., Simmons, Nelson, & Simonsohn, 2012), it is impossible to know how many outcomes were collected but not reported. It is possible that many of these intervention studies, not just those from the Bavelier lab, collected more outcomes than were actually reported in published articles. That sort of underreporting could account for the observed publication bias within the studies by labs other than the Bavelier lab as well.

General Discussion

BAMTGB acknowledged the possibility of publication bias but relied on the trim-and-fill adjustment to conclude in favor of significant benefits from action game training. However, trim and fill is inadequate as

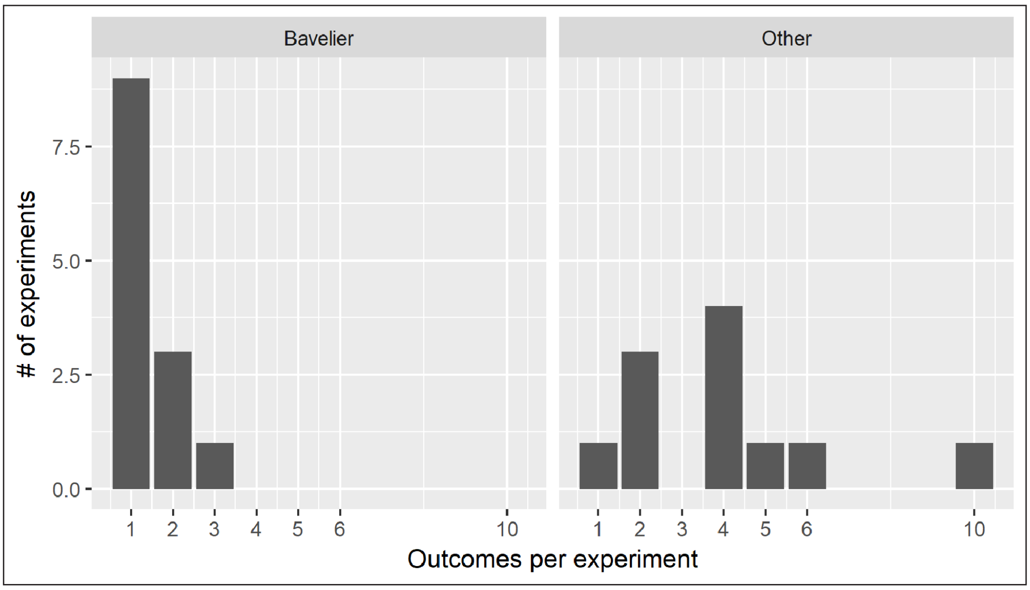

Figure 1: Number of papers (Y axis) in each subset of studies reporting a given number of outcomes (X axis). The subset of papers in BAMTGB by Bavelier and colleagues typically reported fewer outcomes than those by other laboratories.

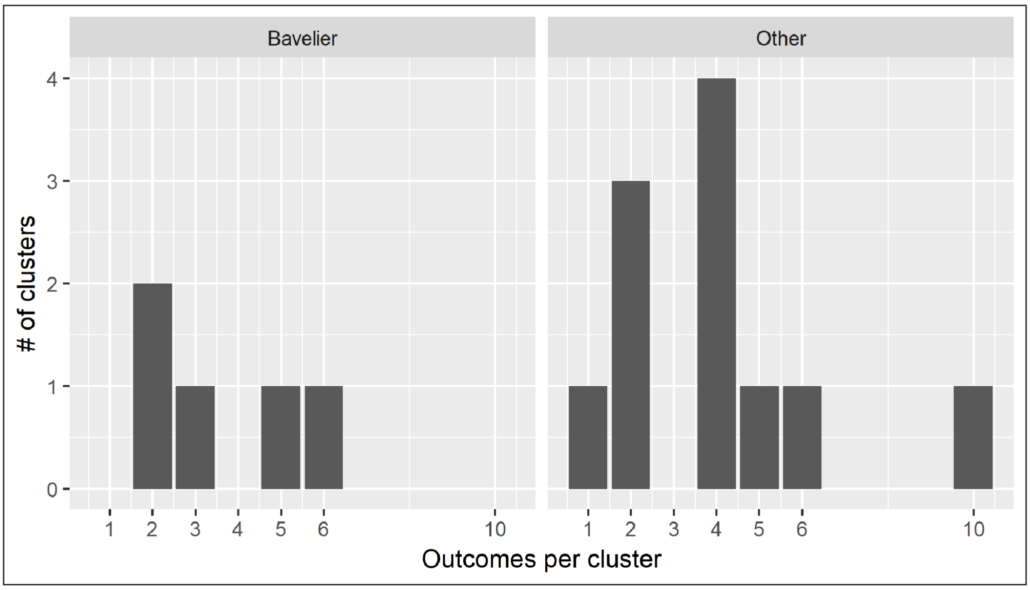

Figure 2: Number of clusters of (partially) dependent samples reporting a given number of outcomes. After clustering together outcomes from overlapping samples that were reported in separate papers, the count of collected outcomes per study is less markedly different.

an adjustment for the substantial publication bias that meta-regression techniques suggest is present (Carter et al., 2019). We applied additional adjustments for publication bias and observed meta-analytic results consistent with a small effect or no effect of action game training on intervention outcomes. That conclusion is consistent across multiple adjustment methods: 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.

Many of the largest effect sizes in published evidence for benefits from action-game training come from papers authored by the Bavelier laboratory. In our analysis, the strong lab effect reported by BAMTGB is consistent with differences in publication bias and is not explained by differences in training duration. An earlier draft of the published erratum for the BAMTGB paper documented how various papers from the Bavelier lab reported outcomes from overlapping sets of participants. Unfortunately, the published version of the erratum removed that "subject overlap" section. The overlap that we document above, combined with the small number of outcomes reported in each paper, suggest selective publication of results from each intervention. It is unclear how many additional outcome measures, particularly non-significant ones, were collected but not reported.

In addition to the issues of salami slicing, the erratum draft suggests other issues that potentially undermine the evidence provided by these studies. First, data collection apparently used a form of rolling recruitment over different time periods, with different outcome measures tested during these periods (Bediou et al., 2018b). This form of study design involves a degree of flexibility that makes it difficult to evaluate what the published results mean. For example, if the completed tasks differed across subjects, it is possible that members of the treatment and control conditions performed different sets of tasks. If so, then the effect of the intervention on any individual outcome measure is confounded by differences in the other tasks completed by those participants. Failure to disclose that recruitment procedure and differences between the tasks completed by different participants could mislead readers and reviewers about the intervention (i.e., participants were not randomly assigned to groups that differed only in the intervention they completed). According to best practices for intervention design, the tasks and outcomes should be identical between the intervention and control group, and the target sample size should be predetermined.

An accurate estimate of the benefits of action video games requires a more complete accounting of the collected samples and outcomes. Future studies should fully report all outcome measures and explicitly describe any overlap with previous studies. For already published research, we hope that the BAMTGB 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 intervention studies 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? We also hope other authors will document the results for any unreported outcome measures from their own intervention studies. Without such an accounting, the statistical inferences reported in the published papers are effectively uninterpretable.

Summary

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 in 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.

Data Accessibility Statement

Data and code are available at https://osf.io/dhejx/.

Notes

[1] 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.

[2] (Retrieved May 22, 2018; https://osf.io/w8xcd/download?version=1&displayName=MA_Intervention_erratum-2018-05-22T17%3A54%3A11.403680%2B00%3A00.pdf).

Additional File

The additional file for this article can be found as follows:

- **Supplement Text S1.** Sensitivity analysis using other, less appropriate combinations of bias adjustment and study-level aggregation methods. DOI: https://doi.org/10.1525/collabra.231.s1

Acknowledgements

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.

Funding Information

GS is a JSPS International Research Fellow (Award #17F17313).

Competing Interests

The authors have no competing interests to declare.

References

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*(1), 77–110. DOI: https://doi.org/10.1037/bul0000130

Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Green, C. S., & **Bavelier, D.** (2018b). Meta-analysis of the impact of action video games on cognition: Intervention studies. Retrieved from https://osf.io/w8xcd/download?version=1&displayName=MA_Intervention_erratum-2018-05-22T17%3A54%3A11.403680%2B00%3A00.pdf

Bejjanki, V. R., Zhang, R., Li, R., Pouget, A., Green, C. S., Lu, Z. L., & **Bavelier, D.** (2014). Action video game play facilitates the development of better perceptual templates. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(47), 16961–16966. DOI: https://doi.org/10.1073/pnas.1417056111

Boot, W. R., Blakely, D. P., & **Simons, D. J.** (2011). Do Action Video Games Improve Perception and Cognition? *Frontiers in Psychology*, *2*, 226. DOI: https://doi.org/10.3389/fpsyg.2011.00226

Carter, E. C., & **McCullough, M. E.** (2014). Publication bias and the limited strength model of self-control: has the evidence for ego depletion been overestimated? *Frontiers in Psychology*, *5*, 823. DOI: https://doi.org/10.3389/fpsyg.2014.00823

Carter, E. C., Schönbrodt, F., Gervais, W. M., & **Hilgard, J.** (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. DOI: https://doi.org/10.1177/2515245919847196

Green, C. S., & **Bavelier, D.** (2003). Action video game modifies visual selective attention. *Nature*, *423*(6939), 534–537. DOI: https://doi.org/10.1038/nature01647

Green, C. S., Pouget, A., & **Bavelier, D.** (2010). Improved Probabilistic Inference as a General Learning Mechanism with Action Video Games. *Current Biology*, *20*(17), 1573–1579. DOI: 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*(3), 984–994. DOI: https://doi.org/10.1016/j.chb.2011.12.020

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Zwienenberg, M., et al. (2016). A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspectives on Psychological Science*, *11*(4), 546–573. DOI: https://doi.org/10.1177/1745691616652873

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

Jobe, J. B., Smith, D. M., Ball, K., Tennstedt, S. L., Marsiske, M., Willis, S. L., Kleinman, K., et al. (2001). Active: A cognitive intervention trial to promote independence in older adults. *Controlled Clinical Trials*, *22*(4), 453–479. DOI: https://doi.org/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*(5), 549–551. DOI: https://doi.org/10.1038/nn.2296

Li, R., Polat, U., Scalzo, F., & **Bavelier, D.** (2010). Reducing backward masking through action game training. *Journal of Vision*, *10*(14), 33–33. DOI: https://doi.org/10.1167/10.14.33

McShane, B. B., Böckenholt, U., & **Hansen, K. T.** (2016). Adjusting for Publication Bias in Meta-Analysis. *Perspectives on Psychological Science*, *11*(5), 730–749. DOI: https://doi.org/10.1177/1745691616662243

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." *Perspectives on Psychological Science*, *11*(4), 512–534. DOI: https://doi.org/10.1177/1745691616635612

Moreno, S. G., Sutton, A. J., Ades, A., 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*(1), 2. DOI: https://doi.org/10.1186/1471-2288-9-2

Office of Research Integrity. (n.d.). Redundancy, Publication Overlap, and Other Forms of Duplication. Retrieved 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 Directions in Psychological Science*, *26*(6), 515–520. DOI: https://doi.org/10.1177/0963721417712760

Simmons, J. P., Nelson, L. D., & **Simonsohn, U.** (2011). False-Positive Psychology. *Psychological Science*, *22*(11), 1359–1366. DOI: https://doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & **Simonsohn, U.** (2012). A 21 Word Solution. *SSRN Electronic Journal*. DOI: https://doi.org/10.2139/ssrn.2160588

Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & **Stine-Morrow, E. A. L.** (2016). Do "Brain-Training" Programs Work? *Psychological Science in the Public Interest*, *17*(3), 103–186. DOI: https://doi.org/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*(6), 666–681. DOI: https://doi.org/10.1177/1745691614553988

Simonsohn, U., Simmons, J., & **Nelson, L.** (2014). Data Colada [30]: Trim-and-Fill is Full of It (bias). Retrieved March 8, 2019, from http://datacolada.org/30

Simonsohn, U., Simmons, J. P., & **Nelson, L. D.** (2015). Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications. *SSRN Electronic Journal*. DOI: https://doi.org/10.2139/ssrn.2694998

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: https://doi.org/10.1037/met0000025

Peer review comments
The author(s) of this paper chose the Open Review option, and the peer review comments are available at: http://doi.org/10.1525/collabra.231.pr

How to cite this article: Hilgard, J., Sala, G., Boot, W. R., & Simons, D. J. (2019). Overestimation of Action-Game Training Effects: Publication Bias and Salami Slicing. *Collabra: Psychology*, 5(1): 30. DOI: https://doi.org/10.1525/collabra.231

Senior Editor: Rolf Zwaan

Editor: Rolf Zwaan

Submitted: 16 January 2019 **Accepted:** 14 June 2019 **Published:** 02 July 2019

UNIVERSITY of CALIFORNIA PRESS Collabra: Psychology

Collabra: Psychology is a peer-reviewed open access journal published by University of California Press.

OPEN ACCESS