Supplement

Text S1

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 estimates the dependency between effect sizes within clusters. 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; BAMTGB present this approach in the original version of the article. 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 application of trim and fill.

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 and the trim-and-fill procedure. 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. This 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 2 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.

Table S1. Other bias-adjusted effect size estimates.

Aggregation	Estimator	All labs	Bavelier lab	Other labs
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]
Averaged	PET	-1.54 [-2.53, -0.54]	-0.26 [-3.23, 2.7]	-1.55 [-2.8, -0.3]
Averaged	PEESE	-0.43 [-0.92, 0.05]	0.41 [-0.93, 1.75]	-0.54 [-1.13, 0.05]
Averaged	SelectionModel	0.25 [0.05, 0.46]	0.72 [0.03, 1.42]	0.21 [-0.02, 0.44]
Averaged	P-uniform	0.43 [0.23, 0.63]	0.95 [0.52, 1.38]	0.29 [0.06, 0.51]
As Indep.	RE	0.34 [0.22, 0.46]	0.92 [0.7, 1.14]	0.22 [0.1, 0.34]
As Indep.	PET	-1.69 [-2.2, -1.19]	-0.24 [-1.48, 1]	-1.89 [-2.6, -1.18]
As Indep.	PEESE	-0.54 [-0.77, -0.3]	0.38 [-0.21, 0.97]	-0.75 [-1.07, -0.42]
As Indep.	SelectionModel	0.06 [-0.17, 0.3]	0.77 [0.41, 1.13]	0.02 [-0.11, 0.16]
As Indep.	P-uniform	0.28 [0.2, 0.36]	0.92 [0.7, 1.14]	0.19 [0.1, 0.27]
Cheung-Chan	RE	0.27 [0.1, 0.43]	0.89 [0.43, 1.36]	0.18 [0.04, 0.32]
Cheung-Chan	PET	-0.04 [-0.27, 0.19]	-0.06 [-2.66, 2.54]	0.01 [-0.24, 0.26]
Cheung-Chan	PEESE	0.12 [-0.03, 0.27]	0.48 [-0.75, 1.71]	0.12 [-0.04, 0.27]
Cheung-Chan	SelectionModel	1.37 [0.13, 2.61]	2.72 [1.73, 3.7]	0.34 [-0.07, 0.75]
Cheung-Chan	P-uniform	0.19 [0.1, 0.28]	NA	0.16 [0.07, 0.25]

Note: Recommended adjusted estimates are presented in Table 1. These other adjustments are provided for the curious reader's sensitivity analysis.