A Second Look at Bias in Violent Games Research: A Reanalysis of Anderson et al. (2010)

Joseph Hilgard, Christopher R. Engelhardt, and Jeffrey N. Rouder

University of Missouri

Author Note

Joseph Hilgard, University of Missouri-Columbia. Please direct correspondence regarding this article to Joseph Hilgard. E-mail: jhilgard@gmail.com

Abstract

Violent video games are theorized to be a significant cause of aggressive thoughts, feelings, and behaviors. A meta-analysis by Anderson and colleagues (2010) is thought by some to condense the research literature into robust and incontrovertible evidence that violent video games affect these outcomes in experimental, cross-sectional, and longitudinal research. In this meta-analysis, application of the trim-and-fill technique found minimal evidence of publication bias. However, there are now more sophisticated methods for the detection of, and adjustment for, publication bias. In the present manuscript, we examine previous meta-analytic evidence and apply these modern techniques for adjusting effect sizes in light of publication bias. Our conclusions differ from those of Anderson and colleagues in three salient ways. First, we detect significant publication bias in experimental research. Second, studies selected as being "methodologically stronger" do not find larger effects than other studies, but instead represent a subsample of the studies in which statistical significance was found. After adjusting for bias, there is no difference between the two estimates. Finally, after accounting for publication bias, effects of violent games on aggressive behavior in experimental research are found to be minimal. That said, it is less clear that effects on aggressive affect and aggressive cognition in experiments have been overstated, as our adjustment techniques disagree as to the magnitude of bias in these studies. Furthermore, the cross-sectional literature appears relatively robust and unbiased. We outline possible sources of bias and suggest directions for stronger future experimental research. The results indicate the need for an open, transparent, and pre-registered research process to test the existence of the basic phenomenon.

A Second Look at Bias in Violent Games Research: A Reanalysis of Anderson et al. (2010)

Do violent video games make their players more aggressive? Despite decades of research and hundreds of studies, the basic phenomena remain, at least for some, controversial. For some authors, whom we term the *advocates*, the answer is definitively in the affirmative. For advocates, the effects are large, obvious, robust, and nearly ubiquitous. For others, whom we term the *skeptics*, the research is not as clean nor as obvious as has been presented. Instead, skeptics point to a host of issues including construct validity, null findings, and publication bias as undermining the evidence for violent game effects. In the writings of the skeptics, the evidence for the violent video game effects is not as solid as claimed; in fact, it is paper thin.

The advocates' primary thesis is advanced by a meta-analysis from (**?**). This meta-analysis covers 381 effect-size estimates based on 130,296 participants. The main findings are that in experiments, there are sizable effects of video game violence on aggressive thoughts ($r = .22$), aggressive feelings ($r = .29$), and aggressive behaviors ($r = .21$). Moreover, these effects not limited to experiments but are also found in cross-sectional comparisons and even in longitudinal research designs. **?** and Huesmann (2010, 2014) call the evidence in this corpus of studies "decisive."

Despite this meta-analysis, there are still skeptics of causal effects of violent video games on aggressive outcomes. Perhaps there are two class of critiques: one is about the *evidentiary value* of the results; the second is about the *interpretation* of the results. The evidentiary critiques are that the meta-analyses suffer from known difficulties including unaccounted publication biases and fortuitous selection criteria for inclusion. Added to this are concerns that the studies themselves suffer from questionable research practices including selective reporting of dependent variables and strategic inclusion of moderating covariates. The interpretation critiques are that violent video games differ from nonviolent games in more than violent content. Such differences include overall arousal, exciting gameplay, and increased competition (**??**). Therefore, it may be difficult to ascribe any

effects to violent content rather than to these other differences.

In this paper, we focus solely on the evidentiary critiques, which are presented subsequently in more detail. We ask whether there is solid evidence for the effect of violent video games on aggressive outcomes. We do not address the question of interpretation here, but see [CITE: Elson & Ferguson; replies to same] for discussion. We ask if there is enough evidence for a violent-video-game effect with the explicit limitation that any such effect may or may not reflect the violent content.

Our approach is to reanalyze the meta-analysis of **?**, and we do so for the following reasons: First, there are now new and more effective techniques for addressing the evidentiary critiques. These new techniques, including PET (Precision-effect test **?**), PEESE (Precision-Effect Estimate with Standard Error **?**), and *p*-curve (**?**)[Note: not Tufano paper], provide for better control of publication bias and questionable research practices than those used in **?**. Second, we along with others (cite Ferguson, 2010; Elson & Ferguson, 2014?) remain concerned with the inclusion criterion. Anderson et al. are admirably transparent in defining two working sets of studies—the full set of all studies, and an additional set of studies that meet *best practices* criteria. As noted by Anderson et al., the best-practices set yields somewhat larger effects than the full set. For example, the violent-video-game effect on aggressive affect is $r = .29$ for the best-practices experiments but only $r = .181$ for the full set.

The fact that Anderson used a best-practices set or that the effects are larger with them than the full set is not in itself of concern. What would be of concern, however, is if there is excessive publication bias and questionable research practices in the sets, and in particular, if there is more of these artifacts in the best-practices set than in the full set. To answer these questions, we provide a reanalysis with funnel plots as well as the PET, PEESE, and *p*-curve estimators that correct for such biases. What we find is indeed concerning—not only is there evidence for bias throughout, there is more bias with the best-practices set than the full set.

Concerns About Bias

In recent years, psychology has experienced a crisis of confidence as researchers realize that many published research findings may be false. Using statistical techniques and reporting standards typical of social psychology, researchers have been able to provide experimental evidence for impossible phenomena such as extra-sensory precogition (psi; **?**) and a song that makes its listeners younger (**?**). Critics have pointed out that hypothesis-confirming results appear in the literature much more frequently than would be expected given reasonable estimates of statistical power. It has even been suggested that the current "publish or perish" reward structure of academia encourages capitalization on Type I error, encouraging researchers to publish many studies with poor predictive value rather than publish few studies with substantial predictive power (**?**). In this light, one might expect that there could be bias in violent games research, as there is in so many other disciplines.

There are two specific types of bias we are concerned with in the meta-analytic context. The first, publication bias, is the phenomenon that studies with statistically significant (i.e., $p < .05$) findings are more likely to be submitted and accepted for publication. Publication bias is a problem that contributes to the overestimation of effect sizes and the propagation of Type I error. It is an especially dangerous problem for meta-analysis, as the selective reporting of studies that "work" (i.e., attain significance) leads to an overestimated effect size and may lead to conclusions of statistically and practically significant effects when there are none. The error introduced by publication bias is larger when research studies are comprised of smaller samples and are consequently underpowered. For these small-sample studies, only those that overestimate the effect dramatically are able to reach the threshold of statistical significance. Hence, small studies with large effects are perhaps the most suspect.

The critical question is whether there is evidence for publication bias in the violent video-game literature. Here, there is disagreement in the literature. Craig et al. claim that

there is little evidence for publication bias. Their claim follows from their attempt to account for such bias. They used a trim-and-fill procedure, which we discuss subsequently, to estimate bias-adjusted effect size estimates. This procedure yielded only a small adjustment suggesting minimal degree of publication bias. This claim strikes us as doubtful for two reasons. First, the authors found 16 dissertations which had found nonsigificant results and subsequently gone unpublished, but only one unpublished non-dissertation study. Given that dissertations likely represent a minority of all studies conducted on violent games, one might expect that there are more unpublished studies yet languish in file drawers. Second, the trim-and-fill correction is understood to be not particularly effective, as it corrects for bias when bias is absent and does not correct enough when bias is strong. **?**, in contrast, makes the case that publication bias is a difficult and pertinent problem in the violent-video-game literature and does so by noting that XXX (JOE). In our view, the claim that there is minimal publication bias in violent media seems implausible given the prevalence of publication bias in research in general and in social psychology in particular. On this basis, more detailed consideration of such bias provided by Anderson et al. is warranted.

The other bias of concern is from practices that inflate Type I error and overstate effect sizes in individual studies. These practices go under many names including *questionable research practices*, *researchers' degrees of freedom*, and *p-hacking*. A common example of such practices include so-called optional stopping where the decision to end a study is dependent on whether a significant effect has yet been found. Some articles such as XXX (REFERENCE, Thinking outside the box) have a collection of experiments where each has a p-value just below the .05 criterion, yet the sample sizes of the experiments vary greatly. We cannot help but wonder if these studies were ran until the p-value was significant.

We suspect there are two specific p-hacking mechanisms in the violent-video game literature. The first has to do with strategic use of dependent measures. In this literature,

it is common to collect several dependent measures. For example, a researcher might measure aggressive behavior by studying both the duration and volume of a retaliatory noise burst. There is considerable diversity in the way studies have combined these quantities, and it has been suggested that the diversity reflects the fact that some results attain statistical significance under one combination while other results attain significance under a different combination. Overall, when researchers collect several dependent measures, there exists the possibility that there is some strategic selection among them. A second mechanism goes by the strategic analysis and presentation of subgroups. In some studies, for example, there are factors manipulated independently of game violence. For example, in **?**, study X, participants play a violent or nonviolent game and are then either clearly or ambiguously provoked by a confederate. In their meta-analysis, Anderson et al. include only those participants in the ambiguous-provocation condition. This selective inclusion was applied in both the best-practices and full-sample analyses. In personal correspondence, Anderson tells us "CRAIG ANDERSON QUOTE", but we find this approach risks capitalizing on chance.

Assessing Bias in Meta-Analysis

There are several approaches to assessing bias in meta-analysis, and some of these have been developed since the publication of Anderson et al. We used several of the more recent tests and methods to provide a new perspective on the Anderson et al. meta-analysis.

A common theme of many of these methods is the relationship between effect size and precision (or sample size) in reported studies. Because sample size does not typically cause effect size, an unbiased research literature is expected to have no relationship between effect size and precision. However, such a relationship will be observed if studies must attain statistical significance to be published. Small-sample studies require large observed effect sizes to reach statistical significance, while large-sample studies can reach

statistical significance with smaller observed effect sizes. Thus, in the presence of publication bias, there is an inverse relationship between effect size and precision.

Note that, in some cases, sample size and effect size may be correlated for reasons other than bias. For example, experimental research tends to have smaller samples than correlational research and may reflect different true effect sizes. Alternatively, it may be possible that manipulations and measurements in small samples are more effective than in large samples. To represent these possibilities, a relationship between sample size and effect size is often called "small-study effects" rather than "publication bias." Some of these possibilities can be excluded through practice; conducting separate bias tests for correlational and experimental research can rule out study design as a potential cause of small-study effects.

Funnel plots. A funnel plot summarizes the relationship between effect size and sample size, allowing for visual estimation of small-study effects. In a funnel plot, effect size is plotted on the x-axis and precision on the y-axis. In the absence of small-study effects or heterogeneity, study results will form a symmetrical funnel shape, displaying substantial variance when sampling error is large but narrowing to a precise estimate when sampling error is small. Thus, when research is not contaminated by bias, some small-sample studies are expected to find null or even negative results due to sampling error. The funnel should fill symmetrically. See Figure 1A for an example of a funnel plot of an unbiased research literature.

Such symmetry is not found in funnel plots of research contaminated with publication bias or *p*-hacking. In the case of publication bias, studies are missing from the lower portion of the funnel where results would not reach statistical significance. See Figure 1B for such an asymmetrical funnel plot. Funnel-plot asymmetry can also be caused by flexibility in analysis and report. When samples are collected until a desired *p*-value is attained, published studies will increase in both precision and effect size, moving towards the upper-right edge of the funnel. When subgroups or experimental subgroups are

dropped from report to highlight only a subgroup in which statistical significance was found, studies will lose precision and increase in effect size, moving towards the lower-right edge of the funnel. When outcomes are censored from report to highlight only the significant outcomes, the effect size increases, moving studies to the right of the funnel.

Again, funnel plots have been presented by skeptics (e.g., **?**), but the **?** meta-analysis did not provide any funnel plots. This makes it difficult for readers to appraise the strength of the data, inspect the distribution of study results, and determine whether the naive and trim-and-fill effect size estimates might be influenced by outliers.

Trim and fill. One popular bias-adjustment technique, trim-and-fill (**?**), attempts to detect and adjust for bias through inspection of the number of studies with extreme effect size estimates on either side of the meta-analytic mean estimate. If the funnel plot is asymmetrical, with many more highly-positive effects than null or negative effects, the procedure "trims" off the most extreme study and imputes a hypothetical censored study reflected around the funnel plot's axis of symmetry (e.g., an imputed study with a much smaller or even negative effect size estimate). Studies are trimmed and filled in this manner until the ranks are roughly equal. See Figures 1C and 1D for examples of trim-and-fill adjusted funnel plots of biased and unbiased literatures, respectively.

However intuitive, this is not an especially effective adjustment for bias, as the assumptions of trim-and-fill are unlikely to be met. Studies are not likely to be censored on the basis of the effect size, but rather, on the basis of their statistical significance. Accordingly, it is argued that trim-and-fill does a poor job of providing an adjusted effect size, adjusting too much when there is no bias and adjusting too little when there is bias (**??**). (Indeed, our simulated datasets in Figures 1C and 1D experience both these problems; however, they are single simulation runs and may not represent the long-run behavior of trim-and-fill.)The imputation of additional effect sizes also must be regarded with caution, as it adds information to the dataset that does not necessarily exist (Higgins & Green, Cochrane Handbook for Systematic Reviews of Interventions, March 2011, v5.1.0)

Thus, trim-and-fill is most commonly suggested as a form of sensitivity analysis rather than a serious estimate of the unbiased effect size. When the naive meta-analytic estimate and the trim-and-fill-adjusted estimate differ only slightly, it is suggested that the research is largely unbiased; when the difference is large, it suggests potential research bias. **?** applied trim and fill in their meta-analysis as the only attempt to detect and adjust for small-study effects. Trim-and-fill yielded only slightly-adjusted effect sizes, and so the authors concluded minimal research bias. Some have characterized this as an extensive test for publication bias (**?**, pg. 51) despite the weaknesses of the trim-and-fill procedure and the absence of funnel plots or other tests for bias.

Egger's regression test. Egger's regression test (**?**) is a simple check for bias which inspects the degree and statistical significance of the relationship between sample size and effect size. A significant test statistic suggests that the observed funnel plot would be unusually asymmetrical were the collected literature unbiased. This test is sometimes helpful in reducing the subjectivity in visually inspecting a funnel plot for asymmetry. Figures 1E and 1F show unbiased and biased research literatures with overlaid Egger regression lines. In the case of the unbiased literature, the slope is not statistically significant, but in the case of the biased literature, the slope is statistically significant, indicating the presence of bias.

One weakness of Egger's regression test is that, while it can detect bias, it does not provide a bias-adjusted effect size. The test is also known to have poor statistical power when bias is moderate or studies are few, limiting the strength of conclusions that can be drawn through application of the test (Sterne, Gavaghan, and Egger, 2000).

Egger's regression test has been used repeatedly by skeptics to look for publication bias (e.g., **??**), but was not reported in the **?** meta-analysis. Thus, while Anderson and colleagues argue that their analysis contains minimal publication bias, an Egger's regression test might have found significant bias.

PET-PEESE meta-regression. A promising new tool in the detection of and adjustment for bias is meta-regression. Meta-regression estimates a bias-adjusted effect size by considering the relationship between effect size and precision, then estimating the underlying effect size that would be found with perfect precision. Two meta-regression estimators are the Precision-Effect Test (PET) and Precision-Effect Estimate with Standard Error (PEESE) (**?**).

In PET, a weighted *linear* regression is fit to describe the relationship between effect size and precision, much like the Egger regression test. Unlike Egger's test, however, PET then extrapolates from this regression to estimate what the "true effect" would be in a hypothetical study with perfect precision. When there is minimal bias, there is minimal adjustment (see Figure 2A). When there is no true effect, published studies tend to lie on the boundary between statistical significance and nonsignificance, forming a linear relationship between sample size and precision. Thus, PET performs well at estimating effects when the null hypothesis is roughly true (see Figure 2C). However, when there is a true effect, small studies will be censored by publication bias, but most large studies will find statistical significance and be unaffected by bias. PET will fail to model this nuance and risks underestimating the size of true effects (see Figure 2B).

A second meta-regression estimator, PEESE, is intended to address this problem. PEESE fits a weighted *quadratic* relationship between effect size and precision. The resulting curve models bias as being stronger in the lower part of the funnel but reduced as the studies become better-powered and less subject to bias. Again, in the absence of bias, adjustment is minimal (see Figure 2D). PEESE is less likely than PET to underestimate nonzero effects (Figure 2E), but risks overestimating the size of null effects (Figure 2F).

The PET-PEESE predictor is intended to address the complementary strengths and weaknesses of the two estimators by combining them in a single conditional estimator. First, PET is applied and the significance of its adjusted effect size is inspected. Next, if the estimate is statistically significant, one is advised to infer a true effect and apply

PEESE to estimate its magnitude. Although this hybrid estimator sounds like it would provide the best of both worlds, the statistical power of PET to detect an effect is unknown, and may be quite poor for sample sizes and effect sizes typical of psychology **?**. Given that a nonsignificant test result does not imply the truth of the null hypothesis, we are reluctant to privilege PET over PEESE. Nonetheless, the PET and PEESE estimators have value as probing the extent of small-study effects; of the two estimators, at least one will be quite good. Thus, the present manuscript reports both PET and PEESE estimates for all meta-regressions. Readers are advised that if the null hypothesis is roughly true, PEESE will overestimate the true effect size, but that if the null hypothesis is false, PET will underestimate the true effect size.

This meta-regression technique has been previously applied by **?** to inspect the amount of evidence for "ego depletion," the phenomenon of fatigue in self-control. They found that after adjusting for small-study effects, PET-PEESE suggested an absence of evidence for the phenomenon. The authors therefore recommended a large-sample pre-registered replication effort, now supported by the American Psychological Society as the topic of the third Registered Replication Report (http://www.psychologicalscience.org/index.php/publications/observer/obsonline/aps-announces-third-replication-project.html).

One criticism of the Egger and PET-PEESE metaregression tests is that some effect size estimates have an inherent relationship between precision and effect size that is not caused by research bias. For example, given a single sample size, the precision of Cohen's d increases as the effect size d increases. A similar phenomenon holds for odds ratio. When these effect sizes are used, metaregression techniques risk misidentifying the inherent relationship between precision and effect size for a small-study effect. To avoid this problem, it has been suggested that one instead use precision estimates that are a function of the sample size alone (Peters, Sutton, Jones, Abrams, & Rushton, 2006). In the current report, we use as our effect size estimate Fisher's Z with standard error $\frac{1}{\sqrt{N-3}}$, consistent

with the original analysis of Anderson and colleagues. Because this standard error is not a function of the effect size, we avoid the problem of an inherent relationship between precision and effect size that might otherwise contaminate the metaregression.

***p*-Curve.** Another novel technique for accounting for small-study effects is *p*-curve (**?**). *p*-curve estimates the true effect size by inspecting the distribution of significant *p*-values. When the null hypothesis is true (i.e. $\delta = 0$), the *p*-curve is flat: significant *p*-values are as likely to be between .00 and .01 as they are between .04 and .05. When the null hypothesis is false, the *p*-curve becomes right-skewed such that *p*-values between .00 and .01 are more common than are *p*-values between .04 and .05. The degree of right skew is proportionate to the power of studies to detect an effect such that increasing sample sizes or larger true effect sizes will yield greater degrees of right skew. By considering the *p*-values and sample sizes of significant studies, *p*-curve can be used to generate a maximum-likelihood estimate of the true effect size.

One weakness of *p*-curve is that, in the presence of questionable research practices, an excess of *p*-values will gather just under the $p = .05$ threshold. This results in a flatter *p*-curve than would be found if studies had been reported without *p*-hacking, and thus *p*-curve will underestimate the true effect size in these circumstances. That aside, simulation work suggests that *p*-curve is quite effective at estimating true effect sizes [CITATION NEEDED].

In summary, we will apply a number of meta-analytic techniques for detecting and adjusting for publication bias. Of these, *p*-curve seems the most promising, but the Egger test and meta-regression estimators also add value.

Unpublished Materials

Publication bias, in which journals tend to publish only significant findings, is a chief source of overestimated effect sizes in meta-analysis. Nonsignificant results can be difficult to retrieve for meta-analysis as they often go unpublished and forgotten. However, one

publication format is largely immune to these publication pressures: the doctoral dissertation. Department requirements generally dictate that dissertations be submitted and published in a dissertation database regardless of whether or not that dissertation is later published as a peer-reviewed journal article. Dissertations are typically thorough, reporting all outcomes and manipulations whereas published journal articles may instead highlight only the significant results. Dissertations, then, provide us with a sample of reported studies relatively uncontaminated by publication biases favoring significant results. In our analyses, we highlight unpublished dissertations, their patterns of statistical significance, and how they fared in meeting best-practices criteria.

Method

We perform a reanalysis of **?** meta-analysis using the same data in the original meta-analysis as provided by the study's first author. We augment the trim and fill approach with funnel plots, PET-PEESE analysis, and *p*-curve effect-size estimation. We use the original authors' separation of studies by study design (experimental, cross-sectional, longitudinal) and by study outcome (affect, behavior, cognition, arousal) in our presentation.

Because the data were analyzed using the software "Comprehensive Meta-Analysis" with the intent of testing for moderators, many studies were entered with separate rows for different outcomes or subsamples within studies. However, our current models of publication bias assume that entire studies are censored or re-analysed per their statistical significance; thus, each study should constitute a single observation. Thus, in the event that multiple effect sizes were entered for a particular study (e.g., effects on mean intensity and count of high intensity trials in the CRTT; separate simple effects for men and women), we aggregated these to form a single effect size for the study. For effects representing separate outcomes within a single sample, the outcomes were averaged. For effects representing separate subsamples within a study, the sample sizes were summed and a weighted average

made of the subsample effect sizes. This parallels the behavior of the software used in the original analysis. p-values were calculated via t-test, first dividing Fisher's Z scores by their standard errors to generate a t-value, then using that t-value to get a two-tailed p-value.

JOE, THERE IS A LOT OF DETAILED CRAP HERE. CAN WE BE MORE DIRECT AND LESS DEFENSIVE? IT FEELS LIKE OVER-DEFENDED OR GRANDSTANDING OR SOMETHING. PLEASE BE MORE DIRECT AND LESS TANGENTIAL. NO QUOTES IN METHOD SECTIONS PLEASE.

We then applied the meta-analytic adjustments. PET was performed by fitting a weighted-least-squares regression model predicting effect size as a linear function of the standard error with weights inversely proportional to the square of the standard error. Similarly, PEESE was also applied, predicting effect size as a quadratic function of the standard error and using similar weights. Finally, p-curve effect size estimates were generated using code provided by **?**, entering a t-value and degrees of freedom parameter for each relevant study.

Within the meta-regressions, all effect sizes were converted to Fischer's Z so as to fulfill the regression model's assumptions of normally-distributed effect sizes. Effect sizes are converted back to Pearson r for tables and discussion. Meta-regression estimates were fit with a multiplicative error term. p-curve estimates were similarly converted from Cohen's d to Pearson r for consistency of presentation.

Both p-curve and PET-PEESE are likely to perform poorly when there are few datapoints. Therefore, our analysis is restricted to effects and experimental paradigms with at least ten independent effect sizes. Our code has been made available online at (GITHUB URL) in the case that the reader nevertheless wants to generate estimates for more sparse datasets or explore the impact of our inclusion and exclusion decisions. The data are available upon request from Dr. Anderson.

In addition to our analysis of the full dataset as provided by Anderson and colleagues, we perform leave-one-out sensitivity analyses, removing each datapoint one at a

time and making all adjusted estimates. For each analysis, a supplementary tab-delimited spreadsheet is attached that lists the individual studies and the estimates when they are left out.[1]

Two studies were removed from the meta-analysis in all analyses. First, **?**, study 1 was removed because its entered effect sizes were unusually large for their precision (i.e., effects on aggressive behavior $r = .60$ and aggressive cognition $r = .53$), were highly influential on the meta-regression model, and most importantly could not be found as entered in the **?** dataset by inspection of the original article. [2] Similarly, **?** was removed because the study tested the effects of violent primes on in-game behaviors and not the effects of violent gameplay itself; therefore, it does not provide a relevant test of the hypothesis.

We reproduce estimates from **?** and apply *p*-curve effect size estimation and PET-PEESE metaregression to detect and adjust for small-study effects. Sufficient datapoints were available to re-analyze experimental studies of aggressive affect, aggressive behavior, aggressive cognition, and physiological arousal, as well as cross-sectional studies of aggressive affect, aggressive behavior, and aggressive cognition. Studies are further

[1]Initially, we had attempted a different sensitivity analysis in which we removed datapoints with a Cook's distance of more than 0.5 on the PET regression. In the case that several observations were excessively influential, we performed an iterative procedure, deleting the single most influential observation and checking again for influence until no observations had excessive influence. In practice, this tended to delete all datapoints that did not fit the PET regression well. This seemed to distastefully and unfairly favor the PET model over the available data; therefore, we eschewed this approach.

[2]We asked Dr. Anderson for comment. He replied, "The Japanese team reported additional results for a number of their papers, in those cases in which the initial paper didn't have what was needed. This was true for several other papers as well. For example, if an original paper reported only some composite measure of aggressive personality but had more specific data on physical aggressiveness, we tried to get the more appropriate measure." It seems unlikely to us that found such a large effect would be found on a single most-appropriate measure that would go unreported in favor of a smaller composite effect. However, it is certainly possible. Without recourse to the raw data, we omit this study as an outlier and probable error of data entry. This footnote is provided for the benefit of the reader so that she may judge our decision.

divided to create separate best-practices-only and all-studies estimates per **?** as sample sizes permit.

Results

Results for all performed *p*-curves and meta-regressions are summarized in Table **??**. Funnel plots with overlaid PET-PEESE regression lines and curves are provided in Figure **??**. We note that visual inspection of the funnel plot often reveals clear asymmetry, particularly in those subsets of studies that **?** selected as "best-practices" studies. Below, we discuss these statistics and describe the results of sensitivity analyses.

Egger's regression test

Results of the Egger's regression tests are supplied in Table XXX. The regression test was statistically significant in several subsets of the data: best-practices and full-sample experiments of aggressive affect, best-practices experiments of aggressive behavior, the full sample of cross-sectional studies of aggressive affect, the full sample (but not best-practices subsample) of experiments of physiological arousal, the best-practices subsample and full sample of cross-sectional studies of aggressive behavior, and the best-practices subsample and full sample of cross-sectional studies of aggressive cognition. The best-practices subsample of experiments of aggressive cognition was also very nearly statistically significant ($p = .055$).

These results indicate that small-study effects are likely present in studies of violent game effects. However, they do not indicate how severe the small-study effects are, or what the true effect sizes may be underlying such small-study effects. We pursue these questions in the next section.

Adjusted effect sizes

Results of the *p*-curve and PET-PEESE analyses are supplied in Table XXX alongside naive fixed-effects and random-effects meta-analytic effect size estimates. Again,

our in-progress simulation work suggests that *p*-curve may be the least biased and most efficient of these estimators. However, a weighted combination of several estimators often outperforms any single estimator. Therefore, we suggest that the reader consider all five estimates and apply her own weights in deciding for herself what seems the most likely true effect in each subsample.

Contrary to the conclusions of the original authors' naive estimates, *p*-curve does not think that best-practices studies measure a larger true effect than do not-best-practices studies. In all cases save one, best-practices and not-best-practices studies received similar adjusted estimates; in the case of correlational studies of aggressive behavior, best-practices studies were estimated as measuring a slightly larger effect.

Because PEESE is thought to be an unbiased estimator of true nonzero effects, one might think that the PEESE estimate approximates an upper bound on the true effect size – an estimate that is accurate if there is indeed a nonzero effect. However, in many cases, the *p*-curve estimate exceeds the PEESE estimate.

There is one notable case in which *p*-curve and PET-PEESE seem to agree on the estimate. When inspecting effects on aggressive behavior in experiments, both techniques estimated that the true effects were very small and likely not meaningfully different from zero. Notably, these estimates are highly consistent with some recent reports by the new generation of violent-media researchers (**??**).

Sensitivity analysis

Leave-one-out sensitivity analyses are presented in a supplementary Excel spreadsheet. We summarize the results below.

Aggressive Affect: Experiments. Among experiments of aggressive affect, it was apparent that one study (**?**) had substantial influence over the meta-regression line, having an extremely large effect size estimate measured with modest precision. After removing this study, the small-study effects were still apparent (best practices, $p_{Egger} = .002$; all

studies, $p_{Egger} < .001$), but meta-regression estimates rose such that PET estimated a more sensible null effect rather than a negative effect (best-practices: PET $r = -.01$, PEESE $r = .17$; full sample: PET $r = -.05$, PEESE $r = .08$). p-curve was not influenced much by this exclusion, recommending $r = .13$ for best-practices and $r = .14$ for full sample.

Aggressive Affect: Correlational. Among cross-sectional studies of aggressive affect, it was found that several of the studies had substantial influence over the PET-PEESE model. The most influential of these was **?**; excluding this study caused the PET estimate to fall to nonsignificance and the effect size to be estimated as $r = .05$. Other influential observations (and the estimated effect size after their exclusion) included **?**, study 2, $r = .13$, and **?**, $r = .16$.

Aggressive Behavior: Experiments. Among experimental studies of aggressive behavior, leave-one-out sensitivity analysis did not indicate major influence of any particular study in the best-practices or full samples. At most, exclusion of **?** sometimes raised the estimate a bit, as one might expect given that it is the study with the largest sample and the smallest effect size.

Aggressive behavior: Correlational. Among cross-sectional studies of aggressive behavior, sensitivity analysis indicated that the estimate was largely robust to the inclusion or exclusion of single studies, with r remaining between .25 and .27 for best-practices and between $r = .18$ and $r = .21$ for full-sample.

Aggressive cognition: Experiments. [THIS NEEDS TO BE REWRITTEN BECAUSE I'M TRYING TO GET AWAY FROM THE NHST IN PET-PEESE.] Because the effect was very near significance, sensitivity analysis suggested some rather variable estimates, as the removal of a single study could cause the p-value to cross the significance threshold. For example, exclusion of **?** caused PET to reach significance, leading to a PEESE estimate of $r = .19$. In the other direction, exclusion of **?** caused the effect size estimate to fall to $r = .06$.

Among all studies, p-curve agreed with the original analysis that the effect was

$r = .21$. PET found a significant effect of violent games on aggressive cognitions ($p = .003$) and no significant small-study effects ($p_{Egger} = .111$). PEESE estimated the effect as $r = .18$, again smaller than the naive or trim-and-fill estimates. Leave-one-out analysis did not detect much variability in estimates, with r ranging from .16 to .19.

Aggressive Behavior: Correlational. Among best-practices cross-sectional studies of aggressive cognition, exclusion of **?** caused the estimate to rise to $r = .17$, while exclusion of **?** caused the PEESE estimate to fall to $r = .13$. When **?** was excluded, the PET estimate fell sharply, no longer reaching statistical significance and recommending $r = .06$. In the full sample, sensitivity analyses indicated two particularly influential observations: exclusion of **?** caused the estimate to rise to $r = .15$, whereas exclusion of **?** caused the PET estimate to no longer reach significance, yielding an estimated effect size of just $r = .04$.

Physiological Arousal: Experiments. In the best-practices subsample, results were highly sensitive to the inclusion or exclusion of single studies, as might be expected of the small number of observations: estimates varied from $r = .08$ to $r = .27$. In the full sample, sensitivity analysis revealed minimal influence from individual studies, with the estimated effect ranging from $r = -.02$ to $r = .02$. Again, p-curve estimates were very different, suggesting an effect *larger* than that of naive meta-analysis, $r = .27$.

Unpublished dissertations

Funnel plots highlighting the unpublished dissertations are provided in Figure YYY. As one might expect given publication bias, the unpublished dissertations generally populate the left side of the funnel plot.

We applied chi-square tests to examine two relationships: first, the relationship between statistical significance and publication status, and second, the relationship between publication status and selection as meeting best-practices criteria. Frequencies are given in Table XXX. The liberal counts assume independence of each entered effect size,

while the conservative counts aggregate all effect sizes within each study.

Chi-square tests were highly significant for all tests. The relationship between statistical significance and publication status was highly significant such that unpublished dissertations were much less likely to have found statistical significance than published studies (liberal and conservative tests, $p < .001$). Similarly, the relationship between publication status and best-practices inclusion was highly significant such that unpublished dissertations were far less likely to be included as best-practices than published studies (liberal test, $p < .001$; conservative test, $p = .002$). Although we had hoped that the application of best-practices criteria would alleviate bias, recognizing well-performed research regardless of its results, it instead appears to have intensified bias.

Discussion

Our findings differ from those of **?** in three important ways. First, we find evidence of publication bias where the original authors argued bias was minimal. Second, the original meta-analysis claimed that methodologically strong studies found larger effects than did methodologically weak studies. Instead, we find that best-practices studies yield estimates comparable to the full set of studies. Division of studies into best- and not-best-practices exacerbated funnel-plot asymmetry, leading to higher naive estimates but comparable adjusted estimates. Third, the original meta-analysis argued that all outcomes were statistically and practically significant. In our analysis, we find instead that the effect of violent video games on aggressive behavior in experiments is likely very small ($r =$ .05–.10). That said, effects on aggressive affect and aggressive cognition in experimental and cross-sectional research seem stronger and more robust, although p-curve and PET-PEESE often disagree about the strength of the effect.

Currently, we believe that p-curve is the stronger meta-analytic technique. Although PET-PEESE is intuitive, easy to visualize, and draws upon more studies than just the statistically significant ones, the power of PET to detect a true effect is questionable,

particularly in sample sizes typical of social psychology. Thus, PET's significance test does not do much to tell us whether PET or PEESE is the better estimator. Nevertheless, we feel that the PET-PEESE estimates add value by representing possible effect size estimates. Future research will be necessary to know how accurate each estimator is.

Although we believe that effect sizes have been overestimated in research, this is not to say that the true effect sizes are precisely as we estimate. First, if the measures and manipulations used by psychologists are ineffective, there may be a true relationship that is not detected. It is possible that 15-minute gameplay experiments are insufficient to observe and test the effects of violent games. Although brief-session experiments of violent game exposure may not detect substantial effects, it is quite plausible that the accumulated effect of many hours of violent gameplay is relevant and detectable, as reported in longitudinal research efforts (citation needed). Second, *p*-curve will underestimate a true effect in the presence of *p*-hacking. Thus, it is possible that the true effect is substantial but our estimates are biased downwards by *p*-hacking in one or more studies. Third, while we find meta-analytic adjustments for research bias useful, we find prospective meta-analysis still more useful. A transparent and pre-registered collaborative replication effort would be ideal.

On the topic of scientific transparency, we note that the clear and accessible archival of meta-analytic data is a tremendous boon to research transparency. We commend Anderson and colleagues for sharing the data and for responding to questions as to how best reproduce their analyses. We suggest that future meta-analyses routinely include the data, funnel plots (in supplemental materials, if need be), and other supplementary materials (**?**). Meta-analyses that cannot be inspected or reproduced should be regarded with concern.

Having detected bias in the meta-analysis, we turn now to possible causes of said bias.

Selection Bias in Meta-Analysis

We observe some instances of flexible application of the best-practices criteria offered by **?**. Flexible application of the inclusion criteria may have lead to preferential selection of studies with significant results. This selection bias could explain why the best-practices studies had larger naive effect-size estimates but comparable adjusted estimates.

p-curve estimates very similar effect sizes for both best-practices and all-studies samples. Recall that *p*-curve inspects only the studies that attained statistical significance. Inspection of the funnel plots reveals that the studies selected as best-practices are generally those studies attaining statistical significance; therefore, the studies considered by *p*-curve are mostly the same across the two samples. We now discuss specific instances of criteria application that may be responsible for selection bias.

Content validity. The first best-practices criterion is that the violent and nonviolent game must be sufficiently different in violent content. Application of this criterion was not consistent. In some cases, studies were excluded for having nonviolent games that contained very mild cartoon violence, while in others, nonviolent games containing substantial violence were included. For example, comparisons between the violent game *Mortal Kombat* and the nonviolent game *Sonic the Hedgehog* were discarded as not-best practices (e.g., **?**) because "the nonviolent game contained violence" (**?**, supplementary materials). Another study comparing a racing game *Moto Racer* against the violent game *Tekken 2* (**?**) was excluded for similar reasons, but we were not able to find any violent content in *Moto Racer*. (At worst, the player can bump into another driver in such a way that both drivers fall off their bikes; neither driver is injured, and the player suffers a time penalty.)

Meanwhile, other studies involving comparisons between violent and not-entirely-nonviolent games were included. **?** was included although it used the game *Final Fantasy* as a nonviolent game. *Final Fantasy* appears to be as violent, or more violent, than *Sonic the Hedgehog*, so the simultaneous inclusion of this paradigm and

exclusion of the *Sonic the Hedgehog* paradigm indicates inconsistency in the application of this criterion. Similarly, a study by **?** was included as best-practices despite comparing the violent *Grand Theft Auto 3* to the purportedly-nonviolent game *Simpsons Hit and Run*. While lighter in tone and less explicit than *Grand Theft Auto 3*, *Simpsons Hit and Run* nonetheless allows the player to punch other characters, steal cars, and run over pedestrians. This content lead video game ratings boards to assign *Simpsons Hit and Run* a rating as appropriate for teens, not children. Thus, again, the application of this criterion seems to favor the inclusion of significant results and the exclusion of nonsignificant results.

Flexibility in the application of this criterion may have contributed to selection biases, inflating the naive meta-analytic estimate relative to the adjusted estimate. A better approach might be to have manipulations rated by research assistants naive to hypotheses or to study results, or to seek a statistical quantification of the difference in violence between games, such as a Cohen's d describing a manipulation check.

Measurement quality. Selection bias may also have been facilitated by the application of best-practices criterion 5: The outcome measure could reasonably be expected to be influenced by the independent variable if the hypothesis were true. For an example of selection bias, see **?**, study 2. In this study, participants were assigned to play a violent or nonviolent game, then complete a competitive reaction-time task measure of aggressive behavior with either an ambiguously or unambiguously provoking confederate. A significant effect was found amount the 90 subjects assigned to the ambiguous provocation condition ($r = .25$), but not among the 90 subjects assigned to the unambiguous provocation condition ($r = -.03$). These 90 subjects with a nonsignificant effect were dropped from both the best-practices and not-best-practices meta-analyses.

When asked for comment, Anderson said "Only the ambiguous provocation condition was used because we now know that the unambiguous (increasing) provocation version of the task is not as sensitive to a variety of independent variables as is the ambiguous provocation pattern. In other words, the increasing provocation conditions don't meet

Criterion 5." While it is possible that only one form of the task is sensitive to the manipulation, the meta-analysis does not seek to model such fine-grained moderators; at the least, the full sample should have been included in the full-sample meta-analysis. Furthermore, the validity or invalidity of measurements cannot be determined on whether they provide the researcher with the desired $p < .05$ in an experiment. Finally, since a significant effect in either the ambiguous or unambiguous provocation group would be taken as evidence for an effect of violent video games, we are concerned that the selective exclusion of groups for not demonstrating such an effect risks introducing selection bias.

Selection bias may also influence which effect size among those reported was entered into analysis. As a general rule, it seems that **?** attempted to avoid subjectivity in effect size entry by averaging all reported effect sizes together. However, on several instances, effect sizes were not averaged together, but rather the single largest available effect size was selected. Returning again to **?**, study 2, the effect of violent games on the first trial of the CRTT was entered (mean difference = 1.07), but not the reported effect size on the other 24 trials of the CRTT (trials 2-9, mean difference = 0.08; trials 10-17, mean difference = 0.04; trials 18-25, mean difference = 0.19). Again, Anderson and colleagues may think that this first-trial-only measure is the most appropriate measurement, at least for this particular study. We are less certain. Selection of the largest effects risks capitalizing on chance and systematically overestimating the true effect. There may be some flexibility involved in the decision to select one trial from a set of twenty-five, to be reported in only one half of the total sample. As **?** point out, not every study uses 1st-trial-only CRTT behavior as the outcome; perhaps the decision to use this particular outcome is contingent on its statistical significance.

In sum, it seems that the inclusion criteria were not effective in selecting an unbiased subset of best-practices studies. Instead, they may have provided some degrees of freedom with which studies with significant results could be included and studies with nonsignificant results excluded.

Omissions

Some null findings were not entered for analysis. In the course of the experiment reported in **?**, a nonexperimental assessment was also made of the effects of previous violent game exposure on aggressive outcomes. In the manuscript, nonsignificant effects of previous violent game exposure were reported for aggressive affect (study 1; $F(1, 66) = 0.78, r = -.11$), aggressive cognitions (study 2; $F(1, 57) = 0.02, r = .02$), and aggressive behavior (study 3; $F(1, 133) = 0.23, r = -.04$). These nonsignificant results were not entered for analysis.

Unpublished Studies

The (**?**) meta-analysis did make an attempt to collect and analyze unpublished studies (e.g., studies presented in dissertations or book chapters that did not undergo peer review). That the resulting analysis remained biased despite these attempts gives us concern that searching for unpublished studies may not actually alleviate bias in meta-analysis.

This is not a criticism of the original authors' meta-analytic effort. Unpublished results are extremely challenging to gather. There is no public record, so database searches will not find them. Many have not been written up, so researchers may not have summary statistics to share with the meta-analyst. Such projects are often forgotten (sometimes deliberately), so even if the meta-analyst asks researchers for unpublished data, it may not be yielded. Finally, null results are sometimes reanalyzed and massaged until they become positive research findings, again censoring null results from public report.

Our inspection of unpublished dissertations suggests that there may be more unpublished non-dissertation studies than just the two found by Anderson and colleagues. This, in accord with our adjustments for small-study effects, suggests that the naive meta-analytic estimate is overestimated by publication bias, and indicates the need for publication of all competent research, not just the research finding significant effects.

Limitations

There are some limitations to the analyses we present. The meta-analytic adjustments used are novel and their limitations may not yet be fully understood. In informal simulations [cite blog posts], *p*-curve tends to perform well. However, it is hard to understand why *p*-curve would estimate effects of violent games on physiological arousal to be larger than would naive meta-analysis. Perhaps some research projects find large effects on physiological arousal but do not report them, as the findings may be considered "too obvious" for publication. Alternatively, perhaps samples are small enough that estimates have substantial imprecision, or we have violated some assumption of the model.

Similarly, PET-PEESE has its own limitations. Although PET seems to perform well when the null is true, and PEESE seems to perform well when the null is not true, the hybrid PET-PEESE technique has questionable power to detect when the null is not true. Thus, PET and PEESE might be thought of as presenting lower and upper bounds on the effect, respectively, rather than identifying the true effect size.

Another criticism of meta-regression is that small-study effects may be caused by phenomena besides publication bias or *p*-hacking. For example, a small survey might measure aggressive behavior thoroughly, with many questions, whereas a large survey can only afford to spare one or two questions. Similarly, sample sizes in experiments may be smaller, and effect sizes larger, than in cross-sectional surveys. The current report is able to partly address this concern by following the original authors' decision to analyze experimental and cross-sectional research separately. Still, there may be genuine theoretical and methodological reasons that larger studies find smaller effects than do smaller studies.

Ways Forward

Although the analyses we present attempt to account for publication and analytic bias, they do not account for validity. Even these adjusted estimates may still overestimate the true effect size due to the influence of confounds. Although it is often claimed that the

observed effects are due to violent content alone (**?**, e.g.,), the evidence for this claim is sometimes weak. Pilot studies are often used to argue that a violent and nonviolent game are equivalent in all other dimensions, but sample sizes are often too small to support this claim (**?**). Application of confounds in analysis of covariance is a more promising approach, but this is also sometimes controversial (**?**). When covariates are measured with error (e.g., with single-item Likert measures), substantial residual variance may be left behind and mistaken for variance associated with violence. Thus, insofar as effects remain after adjustment for small-study effects, they may still be contaminated to some degree by confounds. For these reasons, we favor modified-game paradigms for experimental research (Elson & Quandt, 2014; see Engelhardt, Hilgard, & Bartholow, 2015, Engelhardt et al., 2015, Elson et al., 2015, Kneer, Elson, & Knapp, 2015), which manipulate violent content while preserving the content of gameplay (rules, controls, level design, etc.). One criticism of these paradigms is that the non-violent conditions may not be perfectly nonviolent. We suggest the strengths and weaknesses of these manipulations be the subject of future discussion and study.

We have abstained from inspection of longitudinal studies as there are not enough data points to permit a good estimate. It is possible and even likely that there are detectable longitudinal effects of many hours of gameplay over time. Nonetheless, researchers conducting longitudinal studies should be careful to maintain a transparent research process and to publish results regardless of their significance lest the longitudinal research literature be found to suffer from similar weaknesses.

One line of thought is that the basic phenomenon is certain and that research should be focused on elaborating on the model by exploring moderators of the effect. This perspective is most strongly enunciated by **?**, p. 62, who writes, “Violent media can and must have some psychological impact on those who experience it, and probably does so via well-understood psychological processes. [. . .] Thus, for me, research in media violence no longer needs to establish whether such media can have a psychological and behavioral

impact, but should instead rigorously examine the boundary conditions for such impacts."

We disagree with this perspective regarding the effects on behavior in experiments. We feel that it is most important to establish the existence of the basic phenomenon before attempting to elaborate on possible moderators. If the effects are indeed so small as we estimate, researchers will be hard-pressed to detect the boundary conditions. If *p*-curve is correct and the true effect size in a well-designed experiment is $r = .07$, then 1257 samples are necessary to achieve 80% one-tailed power. To detect the small moderators that reduce the effect to insignificance may require a staggering amount of data.

At present, researchers may feel that they know a lot about the moderators that influence the effect of violent video games on aggressive behavior, as many studies report significant interactions of violent game content by individual differences such as trait anger or gender. However, the diversity of reported moderators and the paucity of replication of these moderations suggest possible weaknesses in the literature. When many moderators are tested, Type I error rates will rise substantially due number of tests conducted. Furthermore, if the main effect is as small as we estimate here, and if the moderating effects are on a similarly small scale, such tests of the interactions could be woefully underpowered, providing little positive predictive value and mostly generating Type I error. Post-hoc exploratory analyses of moderators are valuable (indeed, we have presented them ourselves in the past), but become hazardous when presented as confirmatory or when patterns of statistical significance are taken to identify the validity or invalidity of the measures.

In sum, the research literature as analyzed by (**?**) seems to contain greater publication bias than their trim-and-fill analyses and conclusions indicated. This is especially true of those studies which were selected as using best practices, as the application of best-practices criteria seemed to be influenced sometimes by the results of the study. Effects in experiments seem to be overestimated, particularly those of violent video game effects on aggressive behavior, which appeared to be very close to zero.

Rather than accept these estimates as the "true" effect sizes, we recommend instead a preregistered collaborative research effort and prospective meta-analysis. In this research effort, preregistration and collaboration will both be indispensable. In the absence of preregistration and collaboration, the two well-defined camps of proponents and skeptics may each find results that support their conclusions and refuse to believe the results of the other camp. If we are to advance the debate over violent game effects, we must do it not by silencing the other party, but by getting each party to sit down together with a disinterested third party, design an experiment, and say in writing for all to see, "I agree that this is the appropriate research design. My theory predicts that the result shall be this; his theory predicts that the result shall be that. Together, let us see who is right, and move on."

References

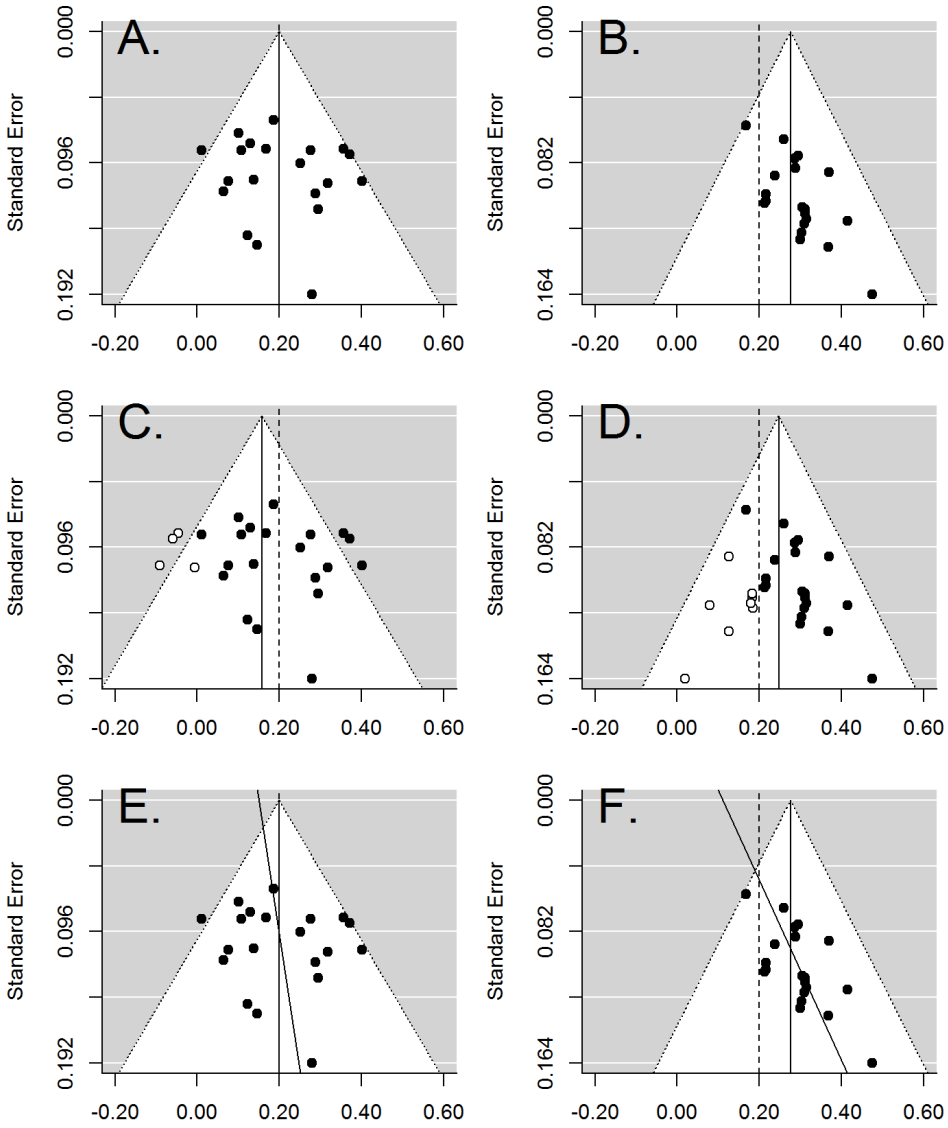

Figure 1. Funnel plots, trim-and-fill, and Egger's test. Effect size Fisher's z is on the x-axis, while standard error of Fisher's z is on the y-axis. The true effect size $z = .2$ is indicated by the dashed line. Panels A and B show funnel plots for unbiased and biased literatures, respectively. The solid line indicates the naive meta-analytic estimate. Panels C and D show the results of trim-and-fill adjustments to these literatures, with the white points representing imputed "filled" studies. The solid line indicates the trim-and-fill-adjusted estimate. Panels E and F show an overlaid Egger's regression line. The slope is statistically significant in F but not in E.

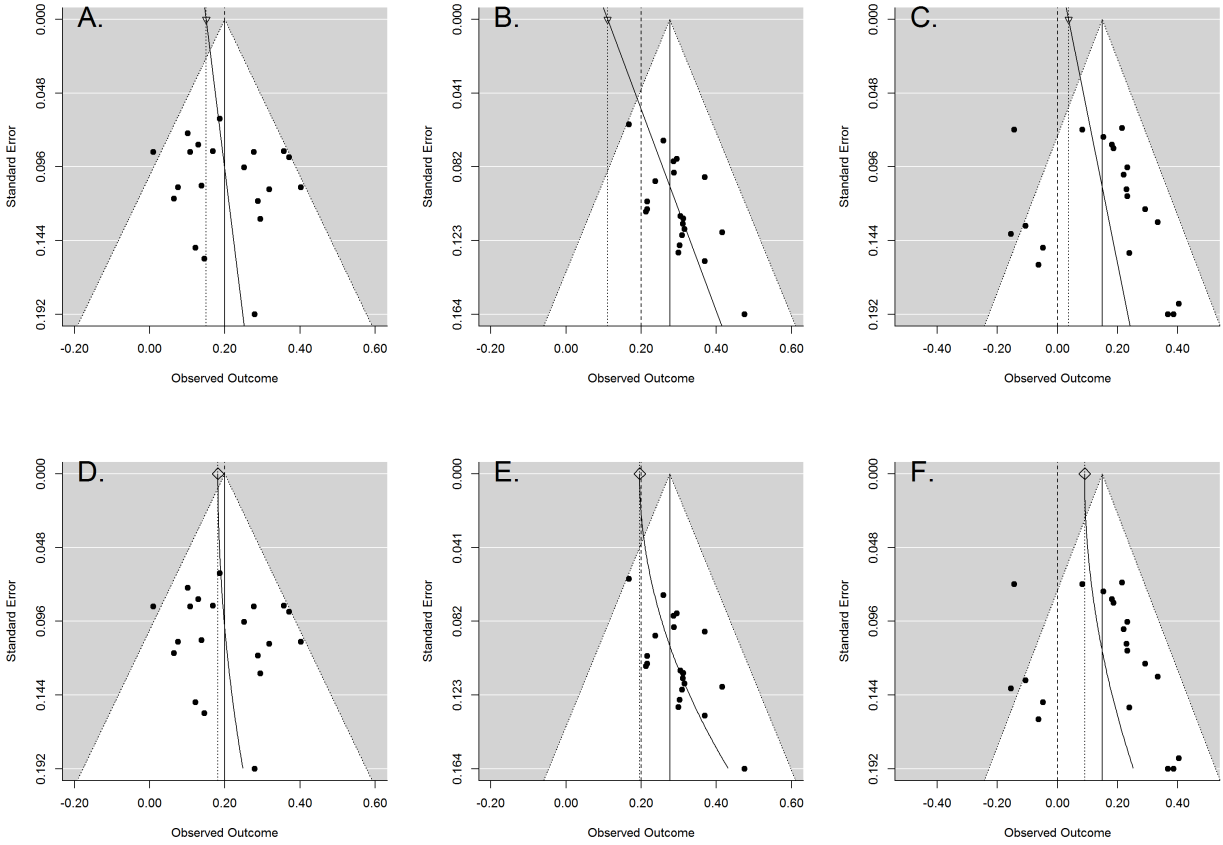

Figure 2. PET and PEESE meta-regression. Again, Fisher's z is on the x-axis, standard error is on the y-axis, and the true effect size is indicated by the dashed line. Bias-adjusted estimates are indicated by the dotted vertical line. Panels A and B indicate the PET technique applied to unbiased and biased literatures of a nonzero effect. PET underestimates the nonzero effect in the presence of bias. Panel C indicates the PET technique applied to a biased literature of a null effect; PET does quite well in estimating the null effect. Panels D and E show PEESE applied to unbiased and biased literatures of a nonzero effect. Panel F shows PEESE applied to a biased literature of a null effect. PEESE does well at estimating the nonzero effect, but overestimates the null effect.