Overestimated Effects of Violent Games on Aggressive Outcomes in 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 that meta-analysis, the authors argued that there is little publication or analytic bias in the literature, an argument supported by their use of the trim-and-fill procedure. However, there are now more sophisticated methods than trim-and-fill for the detection of, and adjustment for, publication bias. In the present manuscript, we re-examine their meta-analysis and apply these new techniques for detecting bias and adjusting effect sizes. Our conclusions differ from those of Anderson and colleagues in three salient ways. First, we detect significant publication bias in experimental research. Second, studies meeting these authors' criteria for methodological quality do not find larger effects than other studies, but instead represent a subsample of studies in which statistical significance was found. After adjusting for bias, there is often little 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, and effects on aggressive affect are much reduced. In contrast, the cross-sectional literature appears relatively robust and unbiased. We outline future directions for stronger experimental research. The results indicate the need for an open, transparent, and pre-registered research process to test the existence of the basic phenomenon.

Overestimated Effects of Violent Games on Aggressive Outcomes in Anderson et al. (2010)

Do violent video games make their players more aggressive? Given the continued popularity of violent video games and their increasing capacity for immersion, even modest effects of violent games could have serious implications for public health. Psychological research provides evidence of such a link, leading professional organizations to issue policy statements concluding harmful effects of violent media (AAP, 2009; APA, 2015). In the view of the professional task forces reviewing the evidence and drafting these statements, the evidence is clear enough, and the hazards certain enough, that the public should be informed and educated of the harmful effects of violent video games.

Despite decades of research and hundreds of studies, the basic phenomena remain, at least for some, controversial. For proponents, the effects are obvious, robust, and nearly ubiquitous. For skeptics, the research is not as clean nor the effects 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 proponents' argument is advanced by a meta-analysis from **?**. This meta-analysis covers 381 effect-size estimates based on 130,296 participants. These estimates were separated into "best-practices" and "not-best-practices" subsets according to whether they met a set of inclusion criteria; the authors emphasize the best-practices subset, but provide analyses of the full sample as a sensitivity analysis. The main findings are that in best-practices experiments, there are statistically and practically significant 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 **?** 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. Skeptics are concerned that the **?** meta-analysis suffers still from purported biases in the publication of studies, the entry of effect sizes into meta-analysis, and the application of the best-practices inclusion criteria (e.g., **?**). Added to this are concerns that the studies themselves suffer from questionable research practices such as the selective report of dependent variables that yield statistical significance (**?**). Skeptics expect that these meta-analytic biases and questionable research practices may overestimate the strength of evidence for, and magnitude of, violent video game effects.

To address this continued skepticism, we re-analyze the meta-analysis of **?**. We feel this re-analysis is necessary for several reasons: First, the topic is important and controversial. Effects of violent video games are hotly debated and have implications for public health and for freedom of expression alike. Second, the **?** meta-analysis is a tremendous volume of work encompassing many studies. We were drawn to the quality and quantity of data. Third, this is purportedly a decisive meta-analysis (see **?**). Good work deserves re-analysis; decisive work *requires* re-analysis. Fourth, there are new and more effective techniques for addressing potential publication bias and questionable research practices. These new techniques, including PET (Precision-Effect Test, **?**), PEESE (Precision-Effect Estimate with Standard Error, **?**), and *p*-curve (**??**), provide for better adjustments for theswe potential artifacts than the method used in **?**.

Concerns About Bias

In many ways, it would be remarkable if violent video game effects were not at least somewhat overestimated, as biases that overestimate effect sizes are common in science. In recent years, psychology has experienced a crisis of confidence as researchers realize that many published research findings may not replicate. In an attempt to replicate the results of 100 psychology studies, only 39 studies yielded the same significant effect (**?**). 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,

suggesting that biases in analysis and report overestimate the strength of effects (see **?**). Such biases can create statistically significant evidence for impossible phenomena. Using statistical techniques and reporting standards typical of social psychology, researchers have been able to provide experimental evidence for extra-sensory precogition (**?**) and a song that makes its listeners younger (**?**). It has even been suggested that the current "publish or perish" reward structure of academia encourages researchers to publish as many Type I errors as possible, which researchers can accomplish by conducting many small, weak studies and using biased analytic techniques (**?**). In this light, one might expect that there could be bias in violent games research, as there is in so many other literatures.

We were concerned about three potential sources of bias in the Anderson et al. meta-analysis. 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. The second, *p-hacking*, is the possibility that researchers increase their Type I error rates in an attempt to find publishable, statistically significant results. The last, *selection bias*, is the application of flexibility in meta-analytic inclusion criteria. We discuss each in turn.

Publication bias. 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 attain significance leads to an overestimated effect size and may lead to unwarranted conclusions of statistically and practically significant effects. 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 as synthesized by **?**. Here there is disagreement. Anderson 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 recommended only a small adjustment, thereby suggesting a minimal degree of publication bias. This claim has two weaknesses. First, the trim-and-fill correction is understood to be somewhat ineffective, as it corrects for bias when bias is absent and does not correct enough when bias is strong (**?**). Other tests for bias have suggested greater evidence of bias (**?**), although this author's execution of the test has been criticized (see **?**). Second, the authors found 16 dissertations which had yielded 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 languishing in file drawers. 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 the possibility of bias in the Anderson et al. meta-analytic dataset is warranted.

***p*-hacking.** Because statistically significant results are easier to publish, particularly in prestigious journals, researchers often strive for statistical significance. Often, this striving leads to the desired statistical significance but also causes an inflated Type I error rate; the obtained result is more likely to be a false positive. Some such practices include data-dependent stopping (i.e., deciding to end data collection when $p < .05$ or continue when $p > .05$), the strategic inclusion or exclusion of outliers depending on their influence on the results, or the analysis of subgroups when results for the whole sample are not found.

P-hacking is, by definition, difficult to detect, as it involves the omission of details that would otherwise influence the interpretation of results. However, it has been argued that there is evidence of *p*-hacking in the quantification and report of certain measures of aggressive behavior. For example, some researchers measure aggressive behavior by

allowing participants to administer a painful burst of noise to another participant. Both the volume and duration of such a noise burst are measured. 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 studies find statistical significance under one combination while other studies find significance under a different combination (**?**). In general, when researchers collect several dependent measures, there exists the possibility that there is some strategic selection among them.

Selection bias. Selection bias may contaminate meta-analysis when the researchers include or exclude studies on the basis of the hypothesis they favor. The application of the best-practices inclusion criteria applied by Anderson et al. was the subject of some controversy. Skeptics argued that the inclusion criteria were applied more liberally to studies with significant results than to studies with nonsignificant results (**?**). If this is the case, then the best-practices subset may find larger effects not due to stronger methodology, but because of greater overestimation through selection bias.

Assessing Bias in Meta-Analysis

There are several approaches to assessing the aforementioned biases in meta-analysis. Some of these rely on statistical procedures, several of which have been developed since the publication of **?**. We used these tests and methods to provide further analysis of the Anderson et al. meta-analysis. Additionally, we looked at the corpus of dissertations not published in journals and considered how their estimates differed from other collected research.

Statistical Procedures

A common theme in many statistical tests for meta-analytic bias 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 studies tend to have smaller samples than correlational studies, and each paradigm may reflect different underlying 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. Funnel plots are without doubt the most important graphical summary of the quality of a meta-analysis. The relationship between effect size and sample size is plotted, 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. Because of this sampling error, some small-sample studies are expected to find null or even negative results even when the underlying effect is positive, so long as there is not bias. See Figure **??**A for an example of a funnel plot of a simulated 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 **??**B for such an asymmetrical funnel plot, generated by simulating a biased research literature.

Publication bias is not the only potential cause of this asymmetry; it 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.

Although funnel plots provide a critical graphical representation of bias, they are, unfortunately, omitted in **?**. This makes it difficult for readers to appraise the strength of the data, inspect the distribution of study results, identify possible mis-entered values, and determine whether the naïve (that is, unadjusted) and trim-and-fill effect size estimates might be influenced by outliers. We provide funnel plots in this report. To foreshadow, our funnel plots show problematic skew in all cases.

One of the critical issues in meta-analysis is what may be learned in the presence of bias. The most charitable position is that researchers may assess the degree of bias provide needed corrections to meta-analytic estimates of effect size (JOE, CITATIONS). Yet, it is our observation that the statistical properties of these corrections—their efficiency and bias in realistically sized samples as well as their robustness to violations of critical assumptions—is certainly suspect in some cases and poorly understood in others. We provide a review of some of these bias-detection-and-correction methods, including a few newer ones, and also note difficulties where they occur.

Egger's regression test. Egger's regression test (**?**) is a simple check for bias that 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 if the collected literature were unbiased. This test is sometimes helpful in reducing the subjectivity in visually inspecting a funnel plot for asymmetry.

Figures **??**E and **??**F show unbiased and biased simulated research literatures with overlaid Egger regression lines. The unbiased literature does not have a significant slope, but the biased literature does.

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. Although Anderson and colleagues argue that their analysis contains minimal publication bias, an Egger's regression test might have found significant bias.

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, 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 **??**C and **??**D for examples of trim-and-fill adjusted funnel plots of simulated 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 **??**C and **??**D 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, 2011).

For these reasons, 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 naïve 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 the trim-and-fill procedure 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. In our opinion, a conclusive test for bias requires more thorough testing than trim-and-fill alone can provide (c.f., **?**).

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 effect would be in a hypothetical study with perfect precision. When there is minimal bias, there is minimal adjustment (see the simulation in Figure **??**A). When there is no underlying 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 underlying effect is approximately zero (see the simulation in Figure **??**C). PET performs less well when there is some effect.

When there is an underlying 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 nonzero effects (see the simulation in Figure **??**B).

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 censoring. Again, in the absence of bias, adjustment is minimal (see the simulation in Figure **??**D). PEESE is less likely than PET to underestimate nonzero effects (Figure **??**E), but risks overestimating the size of null effects (Figure **??**F).

Because PET underestimates nonzero effects and PEESE overestimates null effects, sometimes PET and PEESE are combined as a two-step conditional PET-PEESE procedure. If PET detects a significant effect, the PEESE estimate is used; if PET does not detect a significant effect, the PET estimate is used. Although this approach would seem to make use of the estimators' complementary strengths and weaknesses, this approach may be exceedingly conservative, as PET has questionable statistical power for the detection of effects. When PET's power is poor, conditional PET-PEESE tends to underestimate effects, as only PET is ever applied. For this reason, we report both PET and PEESE. When the PET estimate is significant, the PEESE estimate should be favored, but when it is not significant, one should not necessarily favor PET over PEESE, as non-significant results do not guarantee the truth of the null hypothesis.

These meta-regression techniques have 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 (APS, 2014).

***P*-Curve.** Another novel technique for accounting for small-study effects is *p*-curve (**??**). *p*-curve estimates the underlying 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; larger sample sizes or effects 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. Aside from this weakness, simulation work suggests that *p*-curve is quite effective at estimating true effect sizes (**??**). Another weakness of *p*-curve is that studies with *p*-values above the .05 criterion are not informative resulting in a substantial loss of information. Consequently, the approach is inefficient for estimating small effect sizes.

JOE, SOMETHING ON JOACHIM'S APPROACH. Given this state-of-the-field, our analysis will consists of two main questions. First, are there bias effects in the data? The presence or absence of these effects will be assessed informally by inspection of funnel plots and more formally by the Egger test. Second, if there is bias, what are the appropriate corrections? We will apply *P*-curve, PET, and PEESE to estimate bias-corrected effect sizes. The answer to this second question is necessarily tentative because the statistical properties of these adjustments are only coarsely known.

Unpublished Dissertations

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. Another advantage of dissertations is that they 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 examine unpublished dissertations and their patterns of statistical significance.

Method

We perform a reanalysis of the **?** meta-analysis using the data as provided by the study's first author. We augment the trim-and-fill approach with funnel plots, PET and PEESE meta-regression, 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. Finally, we perform χ^2 tests to see whether unpublished dissertations are more or less likely to yield statistical significance than other published work.

Aggregation of Rows

We assume that entire studies are censored or re-analyzed per their statistical significance. However, the original data have some studies split across multiple rows in order to test for moderators. For example, one study might have two rows: one for the

simple effect among males, and another for the simple effect among females. Where multiple effects were entered for a single study, we aggregated these to form a single row by summing the sample sizes and making a weighted average of the subsample effect sizes. This parallels the behavior of the software used in the original analysis.

Calculation of *p*-values

Although the original data entry performed by Anderson and colleagues is admirably thorough, the data set given us does not have the necessary statistics for *p*-curve meta-analysis. We calculated *t*-values by dividing values of Fisher's *z* by their standard errors, then used the *t*-value to calculate a two-tailed *p*-value. We do not report a *p*-value disclosure table as recommended by **?**, as the meta-analyzed *p*-values are a function of the data as entered by Anderson et al. and not a direct entry of *p*-values from manuscripts. Note that the *p*-values we enter thereby correspond to the main effect of violent video game exposure as entered by Anderson et al. and not the specific hypothesis tests conducted or reported by the studies' original authors.

Adjusted Estimates

PET and PEESE meta-analytic adjustments were calculated. 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. PEESE was also performed, 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 Fisher's *z* so as to fulfill the meta-regression model's assumptions of normally-distributed errors. All meta-regressions were performed using the 'metafor' package for **R** (Viechtbauer, 2010), using the `rma()` function to fit a weighted model with an additive error term. Effect sizes

are converted back to Pearson r for tables and discussion. P-curve estimates were similarly converted from Cohen's d to Pearson r for consistency of presentation.

PET, PEESE, and p-curve 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 https://github.com/Joe-Hilgard/Anderson-meta 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.

Sensitivity analysis. 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. A supplementary spreadsheet is attached that lists the individual studies and the estimates when they are left out.[1]

Studies Excluded

We removed three studies from meta-analysis due to concerns over relevance and accuracy. First, **?**, study 1 was removed because its entered effect sizes were unusually large for their precision (i.e., aggressive behavior $r = .60$ and aggressive cognition $r = .53$), were highly influential on the meta-regression model, and could not be found as entered in the **?** dataset by inspection of the original article. **?** was removed because the study tested the effects of violent primes on in-game behaviors, not the effects of violent gameplay on aggressive outcomes; therefore, it does not provide a relevant test of the hypothesis. Finally, **?** was removed from analysis. As entered in the Anderson et al. dataset, the effect

[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, so we eschewed this approach.

size was unusually large and significant, $r = 0.57, p = 1.6 \times 10^{-10}$. The cause of this enormous outcome was that the study's manipulation checks were entered as though they were primary study outcomes on aggressive cognitions; again, this is not a relevant hypothesis test.

Subsets Re-analyzed

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

The numbers of studies and overall numbers of participants are provided for each subset in Table **??**.

Results

The two key questions are whether there is small-sample bias effects, and if so, what are the best adjustments in effect sizes.

Detection of Bias

The first question is addressed by inspection of the funnel plots in Figures **??**, **??**, **??**, and **??**. Here we see a clear and dramatic asymmetry indicating biases in many cases. These biases are present even in those subsets of studies that **?** selected as best-practices studies. In some cases, funnel-plot asymmetry is more pronounced in the best-practices subsample than in the full dataset.

Results of the Egger's regression tests are supplied in Table **??**. The regression test for funnel-plot asymmetry was statistically significant in several subsets of the data.

Funnel plots were significantly asymmetrical in all analyses of cross-sectional research: affect, behavior, and cognition, both best-practices and full-sample sets. There was also significant asymmetry in both sets of experiments studying aggressive affect. Notably, the Egger test was not significant in the full sample of experiments of aggressive behavior, but it was in the best-practices subsample, suggesting that the application of best-practices inclusion criteria may have exacerbated funnel-plot asymmetry. On the other hand, the full sample of experiments of physiological arousal was significantly asymmetric, whereas the best-practices subsample was not.

In total, these results indicate that small-study effects are likely present in studies of violent game effects. Such a result is concerning because it indicates that the meta-analytic effect size is biased.

Adjusted Effect Sizes

Results of the *p*-curve, PET, and PEESE analyses are supplied in Table **??** alongside naïve fixed-effects and random-effects estimates. We caution the reader that we do not know the small-sample properties of these estimators and so do not valorize one in particular as being likely to provide the most accurate estimate of the underlying effect. Instead, we consider all estimators and look for convergence among adjusted estimates. *P*-values are given for the PET estimate. When the *p*-value is statistically significant, it is suggested that there is evidence of an effect and the PEESE estimate should be favored instead. Again, we caution the reader that a nonsignificant *p*-value does not guarantee that there is no effect.

The current set of estimators yield larger adjustments for bias than did Anderson et al.'s trim-and-fill esitmators. In experiments of aggressive affect, their analyses yielded no adjustment; ours yield an adjustment of -.14 or more. The resulting effect size is $r = .15$ or lower. In experiments of aggressive behavior, their analyses recommended an adjustment of -.03 to $r = .18$; ours recommend an adjustment of -.06–.16 to $r = .05$–.15. In experiments

of aggressive cognition, their analyses recommended an adjustment of -.02 to $r = .20$; our analyses suggest that $r = .20$ could be accurate, but $r = .16$ is also possible.

Contrary to the conclusions of the original authors' naïve estimates, *p*-curve estimates do not yield larger effects in best-practices studies than in the full sample of studies. Only in cross-sectional studies of aggressive behavior do best-practices studies yield larger effects; in all other cases, best-practices subsets and full samples received similar adjusted estimates.

There are some instances of convergence in our presented estimates. When inspecting effects on aggressive behavior in experiments, both *p*-curve and PET estimated that the underlying effects were so small as to be possibly undetectable in typical sample sizes. Notably, these estimates are highly consistent with some recent reports (**????**). For effects on aggressive affect and cognitions in experiments, *p*-curve and PEESE yielded similar estimates, suggesting that there may be detectable effects despite overestimation through research bias. However, we note that the unusually large effect size estimate and unusually small *p*-value obtained in **?** may cause meta-regression to underestimate the effect and *p*-curve to overestimate the effect. Exclusion of this study in our sensitivity analyses dropped the *p*-curve estimate by -.03 , raised the PET estimate to $r = .00$, and increased the PEESE estimates by a further .03.

Unpublished Dissertations

The funnel plots previously presented suggest the presence of substantial bias in publication or analysis. If this is the case, then unpublished dissertations may be less likely to have found statistical significance.

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

We applied χ^2 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 **??**. The liberal counts assume independence of each entered effect size, while the conservative counts aggregate all effect sizes within each study.

All tests were statistically significant. Unpublished dissertations were much less likely to have found statistical significance than published studies (liberal and conservative tests, $p < .001$). Similarly, unpublished dissertations were far less likely to be included as best-practices than published studies (liberal test, $p < .001$; conservative test, $p = .002$). To the extent that these unpublished dissertations may reflect competent research less influenced by publication pressure, these results may be cause for concern.

Meta-analytic effect size estimates were also drastically reduced within the set of dissertations. For aggressive affect, the estimate fell from $r = .17$ [.14, .21] in the full sample to $r = .00$ [-.10, .09] in unpublished dissertations; for aggressive behavior, the estimate fell from $r = .17$ [.14, .20] in the full sample to $r = .01$ [-.11, .12] in unpublished dissertations; and for aggressive cognitions, the estimate fell from $r = .20$ [.17, .23] in the full sample to $r = .13$ [.02, .24] in unpublished dissertations. These estimates should cause pause—they indicate that in effectively preregistered studies, there is scant evidence for any effect of video game violence on aggression.

Discussion

Our findings differ substantially from those of **?** in three important ways. First, we find strong evidence of publication bias where the original analysis argued minimal bias. Egger's test found significant funnel-plot asymmetry, bias-adjusted estimates were substantially smaller than the naïve estimates, and visual inspection of the funnel plots reveals formidable asymmetry. 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 tended to exacerbate funnel-plot asymmetry, leading to higher naïve 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 smaller than anticipated, and may be so small as to be very challenging to study (r = .05–.15). In fact, in light of the results with unpublished dissertations, we remain open to the possibility that there is no effect whatsoever of video game violence on aggressive behavior in experimental settings. That said, effects on aggressive affect and aggressive cognition in experimental and cross-sectional research seem stronger and more robust, although p-curve, PET, and PEESE estimates are in notable disagreement about the strength of the effect.

Limitations

There are important limitations to the analyses we present. Although we have high confidence in the funnel plots to detect bias, we are less sure about the corrected effect size estimates. They certainly represent the best range of answers we think are available (Rosnow and Rosenthal citations). Nonetheless, the statistical properties of these adjustments are not well understood, and the bias and efficiency of these estimators as well as their robustness to misspecification are not known in any systematic or formal fashion, Moreover, they are each understood to perform poorly under certain conditions: PET underestimates non-null effects, PEESE overestimates null effects, and p-curve underestimates effects when there is p-hacking. This limitation of p-curve is particularly salient given concerns about the flexible analysis of the CRTT; it is possible that the underlying effect is substantial but our estimates are biased downwards by p-hacking in one or more studies. It is in this context that we are excited about advances in the field, especially in Vanderkockohve's recent advances with Bayesian meta-analysis. The presented adjustments, in concert with our funnel plots, nevertheless have value in

indicating biases and difficulties in this research literature.

Another limitation 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.

There are also substantive limitations. We have abstained from inspection of longitudinal studies as there are not enough data points to permit a good estimate. It is likely that there are detectable longitudinal effects of many hours of gameplay over time (e.g., **?**), even if the effects of a brief 15-minute exposure in an experiment are undetectably small. All the same, 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. Our point is chiefly that our understanding of the phenomenon as studied through experimental paradigms is likely overstated. Researchers believe they have well-controlled manipulations yielding robust, unbiased effects. We are concerned that, instead, researchers have poorly-controlled manipulations yielding uncertain effects overstated through research bias.

Finally, although the Anderson et al. (2010) meta-analysis is the most-cited meta-analysis finding evidence of effects of violent video games, it is not the only such meta-analysis. A meta-analysis by **?** finds evidence of violent-game effects by summarizing the research literature published since the Anderson et al. (2010) meta-analysis. Our preliminary inspection of their dataset reveals less pronounced funnel plot asymmetry, although we did have to correct their claim that trim-and-fill suggested the effect on

aggressive outcomes had been *underestimated* by bias. (The corrected manuscript now reports no adjustment suggested by trim-and-fill.) We hope to re-analyze this meta-analysis in the future as well.

Implications

Power. The results suggest that individual experiments studying the effects of violent video games may be badly underpowered. If the effects are indeed so small as we estimate, researchers will be hard-pressed to detect them. For example, if we take the *p*-curve adjusted estimate for aggressive behavior in a well-designed experiments, $r = .07$, then 1257 participants are necessary to achieve 80% one-tailed power.

Moderators and boundary conditions. This poor power would have serious implications for the field's understanding of moderators and boundary conditions of violent game effects on aggressive outcomes. Many studies report significant interactions of violent game content by individual differences such as trait anger or gender. We are concerned that the understanding of such nuance is overstated. If the main effects are so small, tests of moderators are likely to be dramatically *underpowered*. If power is poor, the positive predictive value of significant interactions is minimal; such significant interactions would be more likely to be Type I errors than to reflect correctly rejected null hypotheses.

Furthermore, we suspect that significant moderators are tested and discovered *post-hoc*. We expect that it is not unusual to collect a battery of brief personality measures alongside an experimental manipulation. How these measures are to be applied in analysis may be flexible — perhaps they are applied as possible moderators when a significant main effect is not found. When many moderators are tested, Type I error rates will rise substantially due to the number of tests conducted. Post-hoc exploratory analyses of moderators are valuable (indeed, we have presented them ourselves in the past, **?**), but they should be replicated before taken as fact. One of us has published such an interaction, trait anger × violent game exposure (**?**), and has experienced difficulty in replicating it

(Engelhardt, Hilgard, Clark, & Mazurek, submitted). The diversity of reported moderators and the infrequency of their replication suggest possible weaknesses in the literature.

Unfalsifiable predictions of aggressive affect. Of the outcomes we tested, aggressive affect had the most dramatically asymmetrical funnel plot. We suspect that this asymmetry is caused in part by ambiguities in study design and stimulus selection that make it impossible to report null results on aggressive affect. Consider a hypothetical experiment comparing feelings of frustration caused by a violent and a non-violent game. If the result is significant, this is interpreted as evidence that violent video games cause aggressive feelings. However, if the test is not significant, this is not interpreted as evidence that violent games do not cause aggressive feelings — rather, it is taken as evidence that the games are matched stimuli, differing only in violent content and not in other confounding dimensions. Anderson and colleagues (2010) are explicit about this, saying "Studies based on violent and nonviolent video games that have been preselected to be equally arousing obviously are not appropriate tests of the short-term arousal- and affect-inducing effects of violent video games. Thus, they should be excluded from the analyses designed to test this specific hypothesis. The same is true when comparison games have been preselected to create equivalent affective states. (page 156)" Ambiguity in whether the stimuli were truly selected *a priori* on this basis allows for the accumulation of positive evidence and the dismissal of all contrary evidence. Clearly, this tautological situation threatens the fundamental validity of any assessment.

Ways Forward

Meta-analysis, while exciting and informative, is fraught with difficult limitations. We believe that one productive means of avoiding these limitations in assessing the effects of violent media on aggressiveness is conducting large-scale, collaborative, registered, replication reports. In a registered replication report, collaborators review and edit the proposed methods and measures until all agree that the experiment provides a fair and

effective test of the hypothesis. A sample of predetermined size is collected, and the results are published regardless of their statistical significance. This approach protects against biases caused by conditional stopping, flexible analysis, and publication pressures. (Wangemakers citation here)

We suggest that those planning such a registered report consider the use of a modified-game paradigm (**??????**). In such a paradigm, the researchers take a single video game and edit its code. This allows researchers to manipulate violent content while preserving the content of gameplay (rules, controls, level design, etc.). This would minimize concerns that observed effects of violent games are instead due to confounding differences between stimuli. By comparison, usage of commercially-available games does not allow for such control, and differences in violence are likely to be confounded with other differences in gameplay, difficulty, or competition.

Outside of a registered replication effort, there are many other ways to enhance the quality of violent games research. Researchers should consider conducting and publishing direct replications of each others' studies. Larger sample sizes would increase the evidentiary value of individual studies. Preregistration of sample size, measures, manipulations, and analyses would reduce opportunities for conditional stopping (i.e., collecting more data if $p > .05$), censorship of studies or subgroups that fail to find an effect, and flexibility in the quantification of aggressive outcomes. Finally, the open sharing of data would allow for cross-validation: an interaction found in one experiment could then be tested in another researcher's experiment.

Such data-sharing is doubly important in meta-analysis. 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, and other supplementary materials, and that other researchers be encouraged to inspect and reproduce meta-analyses (**?**). Meta-analyses that cannot be inspected or reproduced should be regarded with concern.

t

Summary

he 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. We cannot bear the thought of another thirty years' stalemate. If researchers are to advance the debate over violent game effects, they must do it not by silencing or disgracing each other, but by getting each group 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; their theory predicts that the result shall be that. Together, let us see who is right, and move on."

Table 1

PET, PEESE, and p-curve adjusted estimates.

			Naïve		Adjusted			
	k	N	Fixed	Random	p-curve	PET	p	PEESE
Aggressive Affect								
Experiment - Best	18	1318	0.289	0.335	0.155	-0.120	0.198	0.143
Experiment - Full	34	2879	0.173	0.217	0.164	-0.112	0.055	0.061
Cross-Section - Best	-	-	-	-	-	-	-	-
Cross-Section - Full	14	9811	0.148	0.164	0.164	0.106	**< .001**	0.137
Aggressive Behavior								
Experiment - Best	23	2413	0.209	0.213	0.071	0.072	0.188	0.150
Experiment - Full	39	3328	0.170	0.171	0.052	0.127	**0.003**	0.151
Cross-Section - Best	21	11615	0.263	0.277	0.267	0.227	**< .001**	0.253
Cross-Section - Full	36	28337	0.201	0.229	0.226	0.152	**< .001**	0.189
Aggressive Cognition								
Experiment - Best	24	2887	0.217	0.222	0.185	0.107	0.086	0.180
Experiment - Full	40	4073.5	0.210	0.216	0.205	0.127	**0.008**	0.164
Cross-Section - Best	16	7221	0.168	0.184	0.172	0.099	**0.001**	0.147
Cross-Section - Full	21	12236	0.160	0.191	0.170	0.063	**0.005**	0.130
Arousal								
Experiment - Best	11	833	0.199	0.210	0.262	0.128	0.227	0.183
Experiment - Full	24	1770	0.139	0.148	0.269	-0.005	0.942	0.085

Note: K = number of studies; N = total N across studies. p is p-value for the significance of the PET estimate. When the PET estimate is significant, it is inferred that the underlying effect is nonzero and PEESE should be favored over PET.

Table 2

Egger's regression test.

Outcome	Setting	Sample	*b*	SE(b)	*p*
Affect	Experiment	Best	3.667	0.780	**< .001**
Affect	Experiment	Full	2.743	0.528	**< .001**
Affect	Cross-Section	Best	-	-	**-**
Affect	Cross-Section	Full	1.264	0.640	**0.048**
Behavior	Experiment	Best	1.537	0.549	**0.005**
Behavior	Experiment	Full	0.451	0.390	0.248
Behavior	Cross-Section	Best	1.117	0.483	**0.021**
Behavior	Cross-Section	Full	1.687	0.330	**< .001**
Cognition	Experiment	Best	1.291	0.674	0.055
Cognition	Experiment	Full	0.773	0.479	0.107
Cognition	Cross-Section	Best	1.618	0.649	**0.013**
Cognition	Cross-Section	Full	2.593	0.539	**< .001**
Arousal	Experiment	Best	0.66	0.905	0.466
Arousal	Experiment	Full	1.292	0.626	**0.039**

Note: One may also consider $\alpha = .007$ as the threshold for significance, based on a combination of Egger's recommendation of $\alpha = .10$ and 13-fold Bonferroni correction.

Table 3

The statistical significance and best-practices coding of unpublished dissertations.

Liberal coding scheme.

	Statistical significance	
Publication format	Yes	No
Unpublished Dissertation	3	31
Other	168	155

	Labeled Best Practices	
Publication format	Yes	No
Unpublished Dissertation	4	30
Other	204	119

Conservative coding scheme.

	Statistical significance		
Publication format	Yes	Mixed	No
Unpublished Dissertation	1	2	16
Other	57	42	36

	Labeled Best Practices	
Publication format	Yes	No
Unpublished Dissertation	3	16
Other	81	63

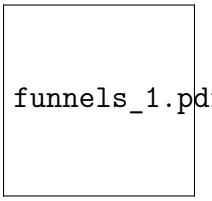

Figure 1. Funnel plots, trim-and-fill, and Egger's test, as demonstrated in simulated data. Effect size Fisher's z is on the x-axis, while standard error of Fisher's z is on the y-axis. The underlying effect size $z = .2$ is indicated by the dashed line. Panels A and B show funnel plots for unbiased and biased literatures, respectively. The solid line indicates the naïve 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.

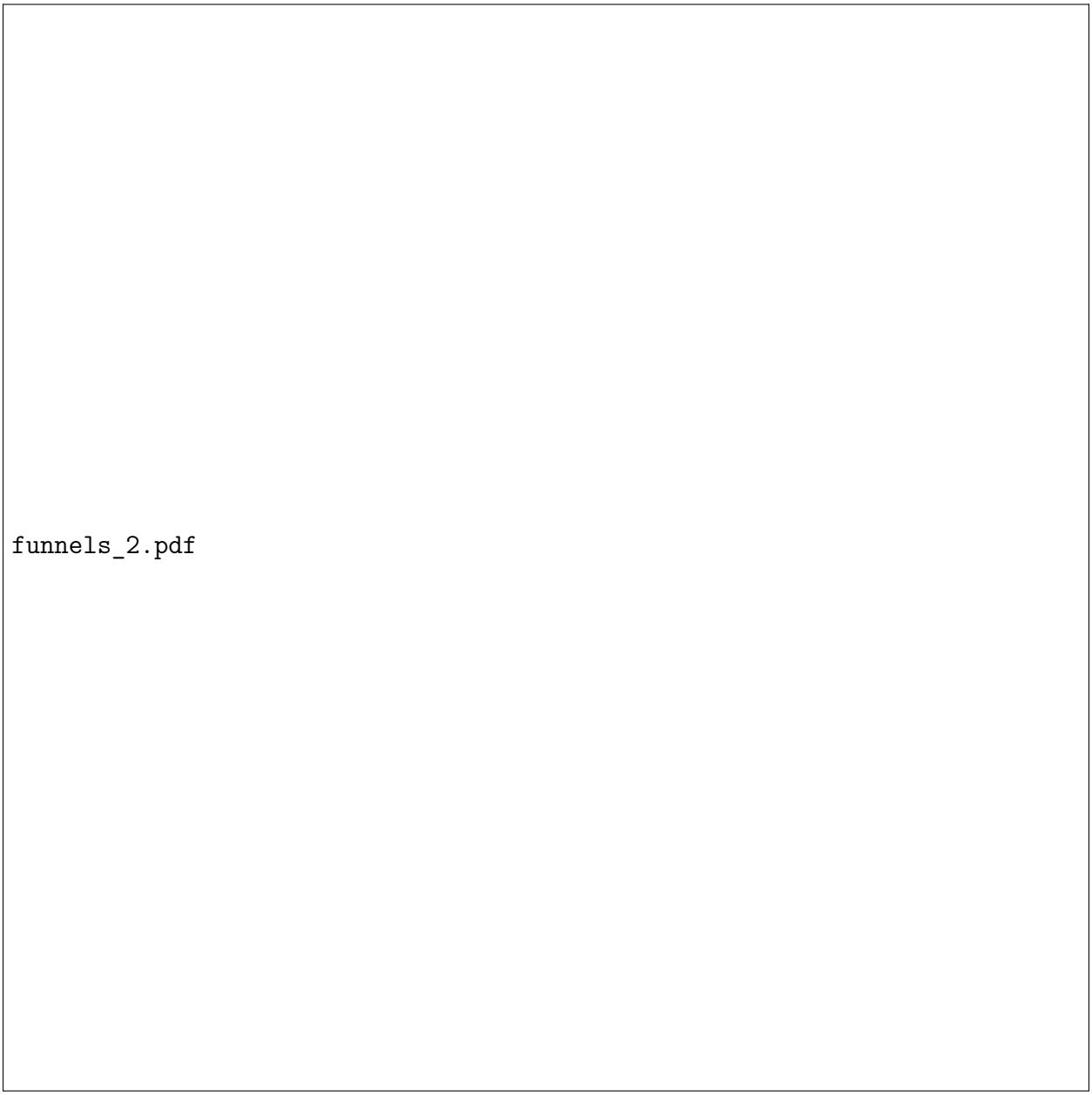

Figure 2. PET and PEESE meta-regression, as demonstrated in simulated data. Again, Fisher's z is on the x-axis, standard error is on the y-axis, and the true effect size is indicated by the dashed line. Filled points represent significant results; hollow points represent nonsignificant results. Bias-adjusted 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

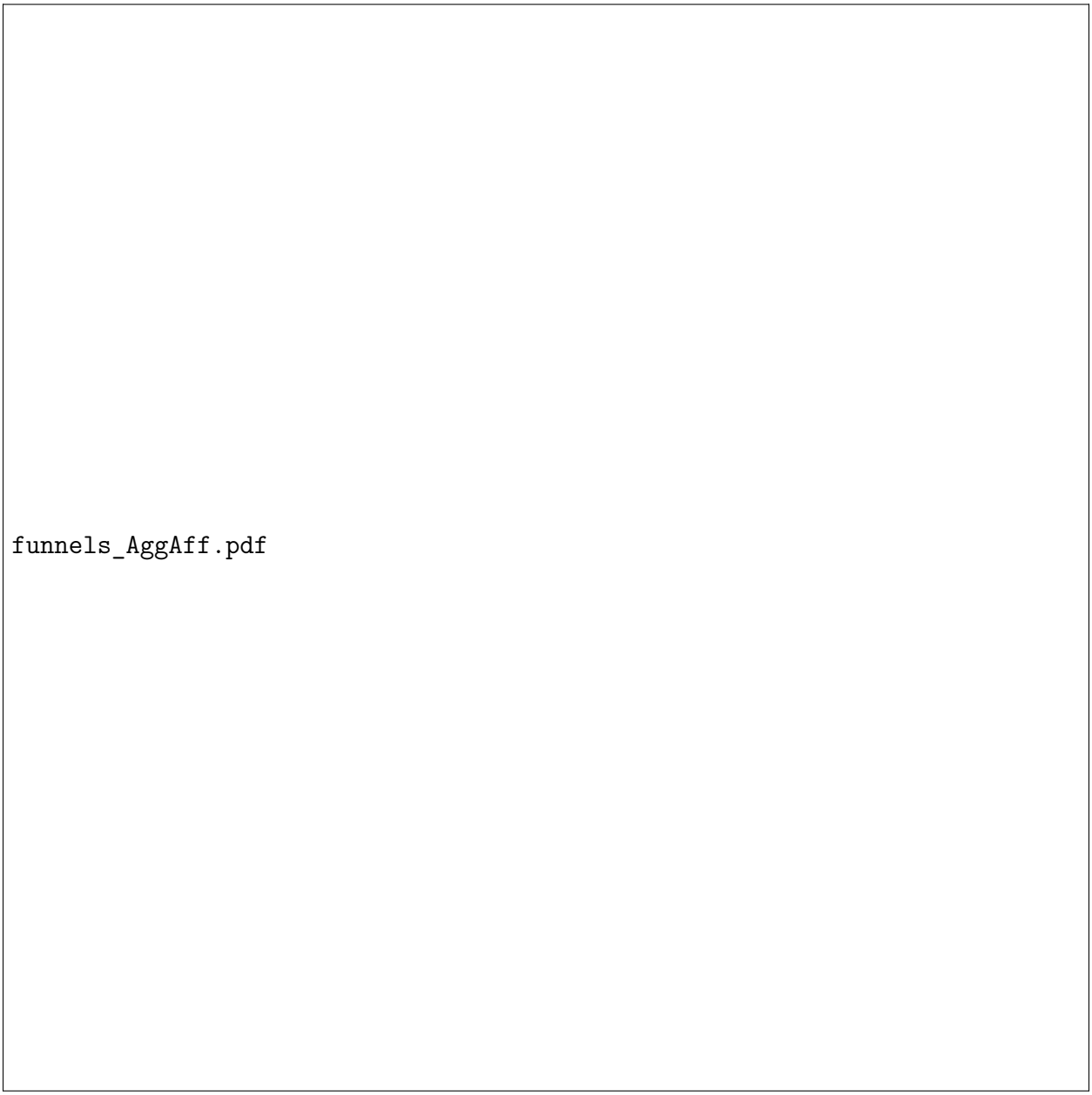

Figure 3. Funnel plot of studies of aggressive affect with overlaid PET and PEESE meta-regression lines. Filled points are statistically significant results; hollow points are nonsignificant results. The circle and diamond points represent the PET and PEESE estimates, respectively. Application of best-practices criteria seems to emphasize statistical significance, and a knot of experiments just reach statistical significance. One experiment (**?**) finds an implausibly large effect, as does one correlational study.

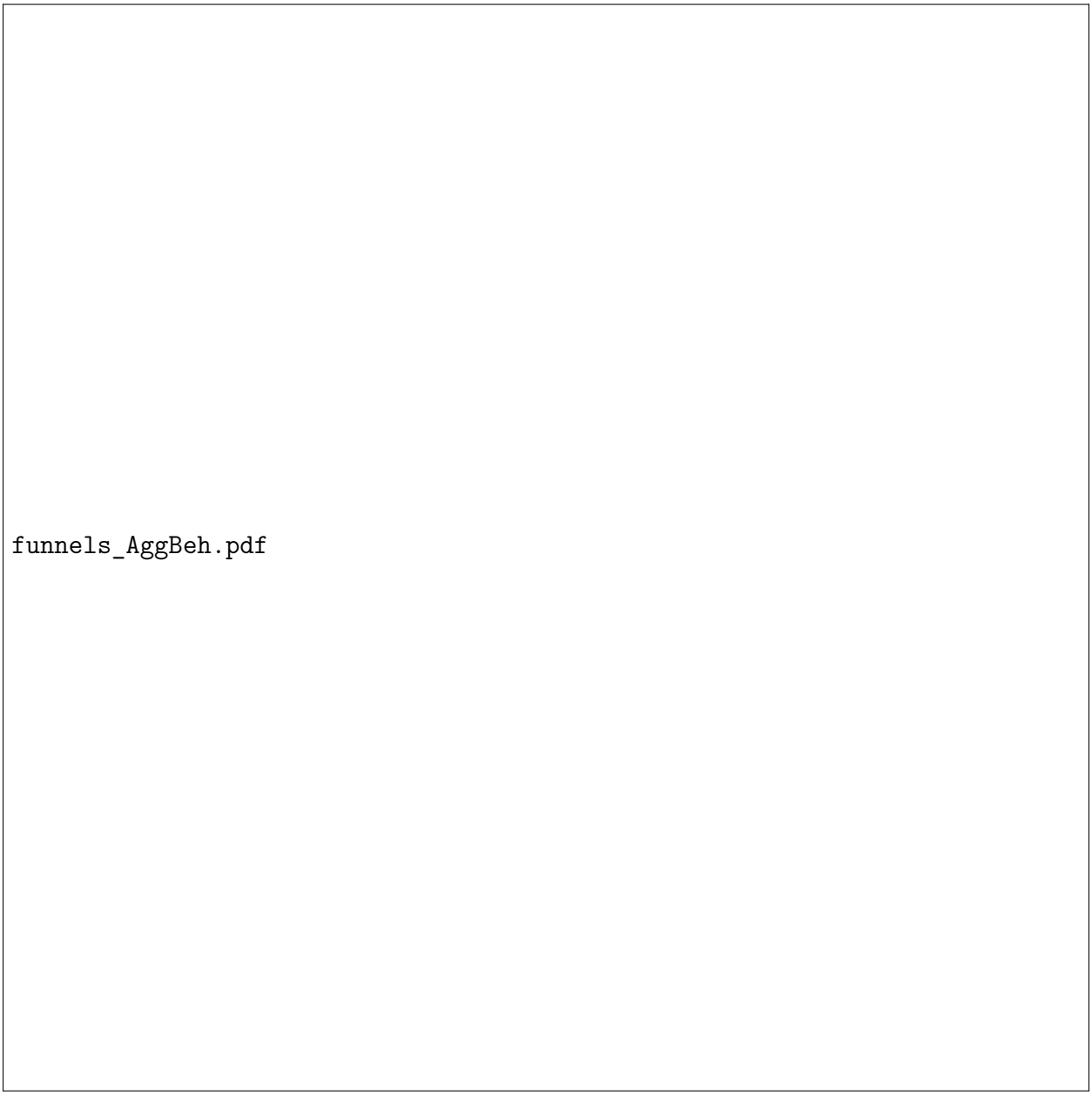

Figure 4. Funnel plot of studies of aggressive behavior with overlaid PET and PEESE meta-regression lines. Filled points are statistically significant results; hollow points are nonsignificant results. The circle and diamond points represent the PET and PEESE estimates, respectively. Application of best-practices criteria seems to emphasize statistical significance, and a knot of experiments just reach statistical significance. Again, application of best-practices criteria favors experiments finding statistical significance.

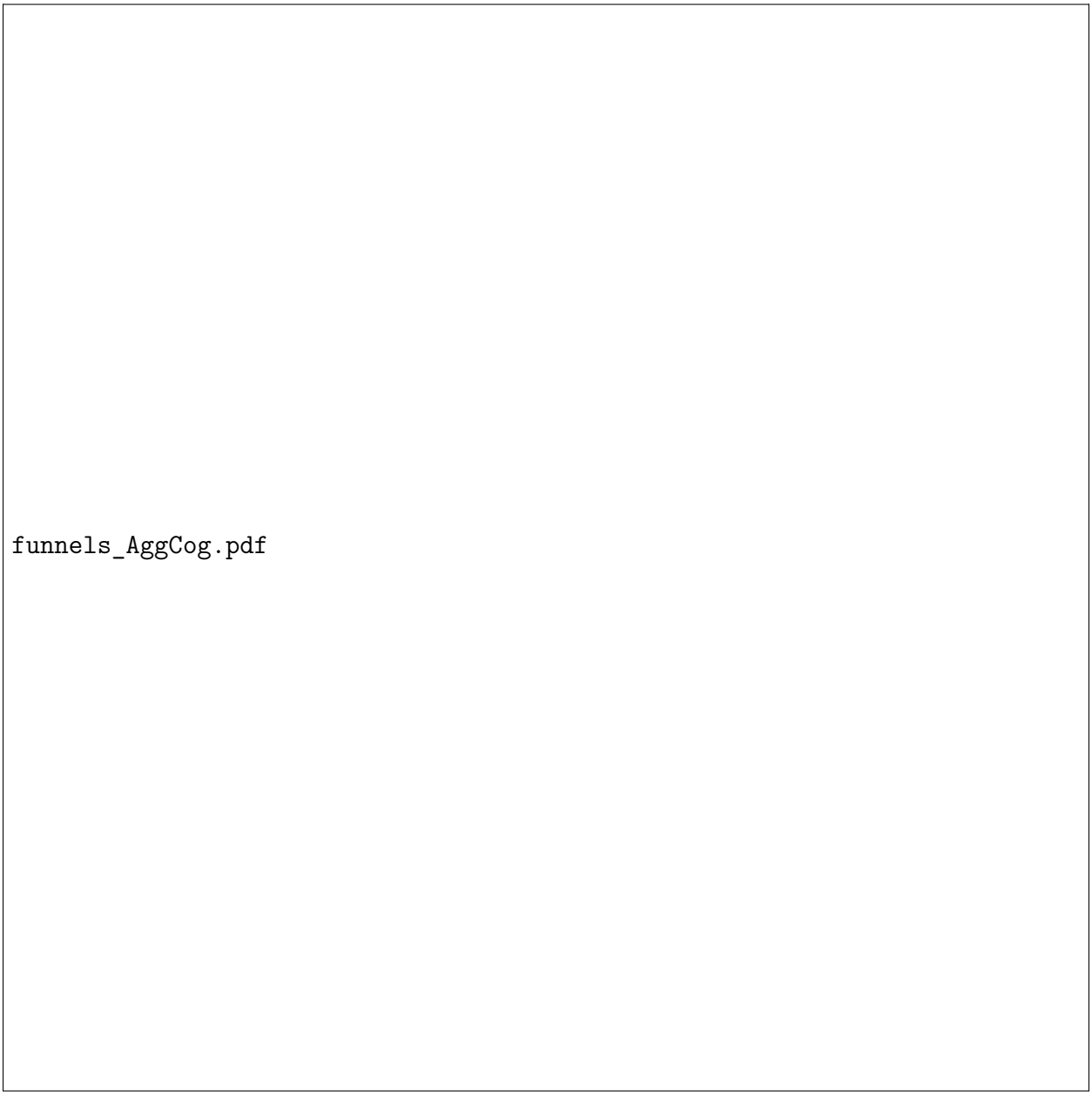

Figure 5. Funnel plot of studies of aggressive cognition with overlaid PET and PEESE meta-regression lines. Filled points are statistically significant results; hollow points are nonsignificant results. The circle and diamond points represent the PET and PEESE estimates, respectively. Results appear heterogeneous, but somewhat less contaminated by bias.

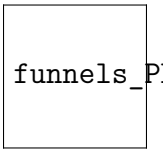

Figure 6. Funnel plot of studies of physiological arousal with overlaid PET and PEESE meta-regression lines. Filled points are statistically significant results; hollow points are nonsignificant results. The circle and diamond points represent the PET and PEESE estimates, respectively. Application of best-practices criteria seems to emphasize statistical significance, and a knot of experiments just reach statistical significance. Results do not appear to be systematically contaminated by bias.

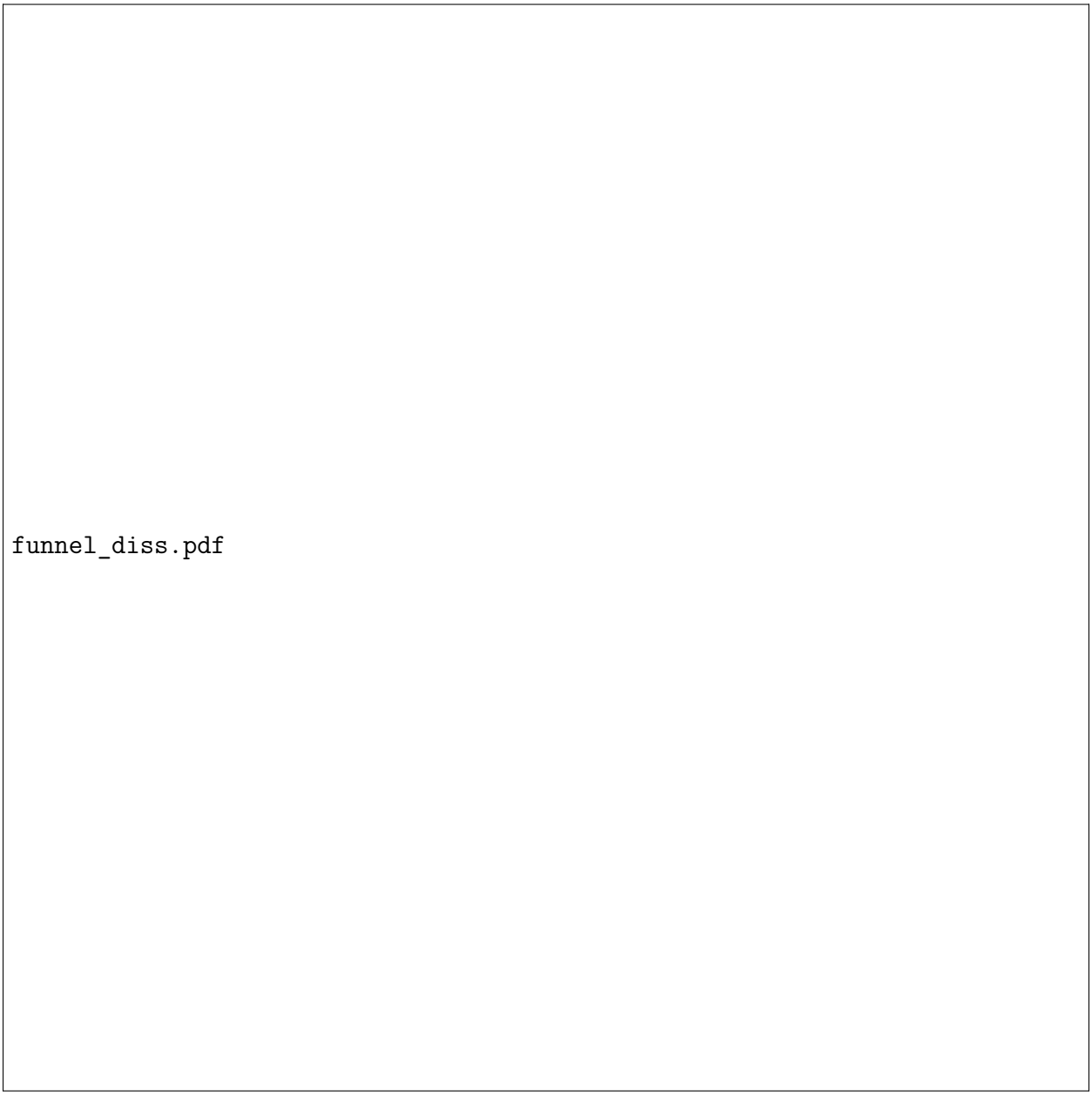

Figure 7. Funnel plots of all experiments of aggressive affect, behavior, and cognition. Dissertations not presented in any further publication format are indicated with Xs, while all other publication styles (e.g., journal articles, book chapters, conference proceedings) are indicated with filled dots. The shaded band represents *p*-values between .05 and .01. Nonsignificant results are less likely to be published, and in the case of experimental studies of affect and of behavior, dissertations suggest substantially smaller effects.