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 new 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 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, and effects on aggressive affect are much reduced. On the other hand, the cross-sectional literature appears relatively robust and unbiased. We outline possible sources of bias and suggest 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.

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 proponents, the effects are large, 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 Anderson et al. (2010). 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. Bushman, Rothstein, and Anderson (2010) and Huesmann (2010) 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. There are two classes 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 purported 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 moderators or covariates. The interpretation critiques are that violent video games differ from nonviolent games in more than violent content. Such purported differences include overall arousal, exciting gameplay, and increased competition (Adachi & Willoughby, 2011; Elson,

Bruer, Van Looy, Kneer, & Quandt, 2013). 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 Elson and Ferguson (2013) and Bushman and Huesmann (2014) 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 effect of violent content in specific.

Our approach is to reanalyze the meta-analysis of Anderson et al. (2010), and we do so for the following 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. Finally, this is purportedly a decisive meta-analysis (see Huesmann, 2010). Good work deserves re-analysis; decisive work *requires* re-analysis.

Our re-analysis offers new insights for the Anderson et al. meta-analysis. First, now there are new and more effective techniques for addressing the evidentiary critiques. These new techniques, including PET (Precision-Effect Test, Stanley & Doucouliagos, 2014), PEESE (Precision-Effect Estimate with Standard Error, Stanley & Doucouliagos, 2014), and p-curve (Simonsohn, Nelson, & Simmons, 2014a, 2014b), provide for better adjustments for publication bias and questionable research practices than those used in Anderson et al. (2010).

Second, some are concerned about the application of the inclusion criteria (Elson & Ferguson, 2013). Anderson et al. are admirably transparent in defining two working sets of studies—the full set of all studies and a subset of studies that meet *best practices* criteria. As noted by Anderson et al., the best-practices subset 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 study results influenced inclusion or exclusion from the best-practices set.

To answer these questions, we provide a reanalysis with funnel plots as well as the PET, PEESE, and *p*-curve estimators that adjust for such biases. We find, to our concern, that not only is there evidence for bias throughout, but 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 not replicate. 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 (Bem, 2011) and a song that makes its listeners younger (Simmons, Nelson, & Simonsohn, 2011). In an attempt to replicate the results of 100 psychology studies, only 39 studies yielded the same significant effect (Open Science Collaboration, 2015). 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 (see Schimmack, 2012). 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 (Bakker, van Dijk, & Wicherts, 2012). 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. 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 strikes us as doubtful for two reasons. First, 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 (Simonsohn et al., 2014b). Ferguson (2007), in contrast, makes the case that publication bias is a problem in the violent-video-game literature through application of Egger's regression test. 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 at the cost of 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.

We suspect there are two specific *p*-hacking mechanisms in the violent-video game literature. The first involves the strategic use of dependent measures. In this literature, it is common to collect several dependent measures. 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 (Elson, Mohseni, Breuer, Scharkow, & Quandt, 2014). In general, when researchers collect several dependent measures, there exists the possibility that there is some strategic selection among them.

A second mechanism is the strategic analysis and presentation of subgroups. For example, in Anderson et al. (2004), study 2, participants played a violent or nonviolent game and were 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, "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." (Personal communication, Nov. 4, 2014). We find this approach risks capitalizing on chance, allowing extra opportunities for an effect to be detected and thereby increasing familywise error rates.

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. The inclusion criteria seemed to be applied more liberally to studies with significant results than to studies with nonsignificant results. [cite Ferguson, 2010? Elson & Ferguson, 2013?] 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, and some of these have been developed since the publication of Anderson et al. (2010). 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 studies tend to have smaller samples than correlational studies, and each paradigm 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. Because of this sampling error, some small-sample studies are expected to find null or even negative results even when the true effect is positive, so long as there is not bias. 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.

Although funnel plots provide a useful graphical representation of bias, they are, unfortunately, omitted in Anderson et al. (2010). 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 them in this report.

Trim and fill. One popular bias-adjustment technique, trim-and-fill (Duval & Tweedie, 2000), 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 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 (Simonsohn et al., 2014b). 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 (Simonsohn et al., 2014b). (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, 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. Anderson et al. (2010) 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. In our opinion, a conclusive test for bias requires more thorough testing than trim-and-fill alone (c.f., Bushman & Huesmann, 2014).

Egger's regression test. Egger's regression test (Egger, 1997) 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 if the collected literature were 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. 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., Ferguson, 2007; Ferguson & Kilburn, 2009), but was not reported in the Anderson et al. (2010) meta-analysis. Although 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) (Stanley & Doucouliagos, 2014).

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). PET performs less well when the null hypothesis is false. 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 censoring. 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).

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, we do 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 Carter and McCullough (2014) 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 as 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 (Simonsohn et al., 2014a, 2014b). *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 (Simonsohn et al., 2014a, 2014b).

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. 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, their patterns of statistical significance, and how they fared in meeting best-practices criteria.

Method

We perform a reanalysis of the Anderson et al. (2010) 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.

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 Fisher's Z scores 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 Simonsohn et al. (2014a), 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.

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 Simonsohn et al. (2014a), 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 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://collaborate.missouri.edu/jhilgard/craig_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. For each analysis, a supplementary tab-delimited 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, Matsuzaki, Watanabe, and Satou (2004, 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 could not be found as entered in the Anderson et al. (2010) dataset by inspection of the original article.[2] Panee and Ballard (2002) was removed

[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.

[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 such a large effect would be found on a single most-appropriate measure and nevertheless 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.

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, Graybill, Kirsch, and Esselman (1985) was removed from analysis. As entered in the Anderson et al. dataset, the effect 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. Participants were asked "Tell me what happened in the video game you played," and "What did you like about the video game you played?" The results of the chi-square tests on these manipulation checks were then averaged together with two non-significant outcomes. While what is meant by "aggressive cognition" is not exactly clear, we do not think this manipulation check provides a relevant test.

Subsets re-analyzed

We reproduce estimates from Anderson et al. (2010) 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 Anderson et al. (2010) as sample sizes permit.

Results

Results for all performed *p*-curves and meta-regressions are summarized in Table 1. Funnel plots with overlaid PET regression lines and PEESE curves are provided in Figures 3, 4, 5, and 6. We note that visual inspection of the funnel plot often reveals clear asymmetry, particularly in those subsets of studies that Anderson et al. (2010) selected as "best-practices" studies. Often, the best-practices subsample seems to preferentially exclude studies from the lower-left side of the figure. Below, we discuss these statistics.

Egger's regression test

Results of the Egger's regression tests are supplied in Table 2. 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, PET, and PEESE analyses are supplied in Table 1 alongside naïve fixed-effects and random-effects 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.

p-values are given for the PET estimate. When the *p*-value is statistically significant, it is suggested that there is a true 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 true effect.

Contrary to the conclusions of the original authors' naïve 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.

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 (Engelhardt, Mazurek, Hilgard, Rouder, & Bartholow, 2015; Hilgard, 2015; Kneer, Elson, & Knapp, in press; Przybylski, Deci, Rigby, & Ryan, 2014).

Unpublished dissertations

Funnel plots highlighting the unpublished dissertations using experimental paradigms are provided in Figure 7. 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 3. 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.

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.

Discussion

Our findings differ substantially from those of Anderson et al. (2010) in three important ways. First, we find strong 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 naïve estimates but comparable meta-regression-adjusted estimates. Similarly, *p*-curve estimated very similar effect sizes for both best-practices and all-studies experiments. Like Anderson's best-practices analyses of experiments, the *p*-curve technique inspects only statistically significant results 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 small ($r = .05$–$.15$). 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.

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 (e.g., Willoughby, Adachi, & Good, 2012). 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. Meta-regression may suffer from similar underestimation. 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, and other supplementary materials, and that other researchers be encouraged to inspect and reproduce meta-analyses (Lakens, Hilgard, & Staaks, in press). 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. These include publication bias, selection bias of entered effect sizes, and selection bias of studies said to meet best-practices criteria.

Publication Bias

Anderson and colleagues did make a commendable 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' specific meta-analytic effort. Rather, it is a reminder that 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, 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; there may not even be null results to gather.

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 naïve meta-analytic estimate is overestimated by publication bias and indicates the need for publication of all competent research regardless of its results.

Selection Bias in Data Entry

Some null findings were not entered for analysis. In Carnagey and Anderson (2005), three experiments were conducted examining the effects of violent game play on aggressive affect, cognitions, and behavior, respectively. Along with these experimental manipulations of violent game exposure, the authors measured participants' history of violent game exposure. In each study, the authors tested whether this was related to two dependent variables: the primary study outcome (affect, cognition, or behavior) and a questionnaire measure of trait physical aggression. Past violent game use was not significantly associated with the primary outcomes , but it was significantly associated with trait physical aggression. The significant association was entered for meta-analysis but the nonsignificant associations were not. Although we are confident that Anderson et al. mean well and view the significant association as a more-appropriate test, we think both effect sizes should

have been entered and averaged for analysis.

Selection Bias in Best-Practices Criteria

We observe some instances of flexible application of the best-practices criteria offered by Anderson et al. (2010). 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 experiments had larger naïve effect-size estimates but comparable adjusted estimates.

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.

Dissertations with nonsignificant results were excluded as failing to meet this criterion despite considerable differences in violent game content. Comparisons between the violent game *Mortal Kombat* and the nonviolent game *Sonic the Hedgehog* were discarded as not-best practices (Cohn, 1995; Hoffman, 1994) because "the nonviolent game contained violence" (Anderson et al., 2010, supplementary materials). Another study comparing a racing game *Moto Racer* against the violent game *Tekken 2* (Brooks, 2000) 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, published research finding significant differences were included as meeting this criterion, even when the differences in violent game content were more subtle. Konijn, Nije Bijvank, and Bushman (2007) 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 Brady and Matthews (2006) was included as best-practices despite comparing the violent *Grand Theft Auto 3* to the purportedly-nonviolent game *Simpsons Hit and Run*. Although lighter in tone and less explicit in content than *Grand Theft Auto 3*, *Simpsons Hit and Run* nonetheless allows the player to punch other characters, steal cars, and run over pedestrians. Video game regulatory boards rated *Simpsons Hit and Run* and *Final Fantasy* ratings as being appropriate for teens, not children.

Flexibility in the application of this criterion may have contributed to selection biases, inflating the naïve meta-analytic estimate relative to the adjusted estimate. A better approach might be to have manipulations rated by research assistants blind to hypotheses or to study results, or to seek a statistical quantification of the difference in violence between games, such as an effect size estimate of a manipulation check measuring violent content.

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 Anderson et al. (2004, 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. The increasing provocation conditions don't meet Criterion 5." At the least, the full sample of ambiguously and unambiguously provoked participants 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 the introduction of selection bias.

Selection bias may also influence which effect size among those reported was entered into analysis. As a general rule, it seems that Anderson et al. (2010) 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 Anderson et al. (2004, 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 Elson et al. (2014) 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.

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. *p*-curve tends to perform well in simulations, but it is hard to understand why *p*-curve would estimate effects of violent games on physiological arousal to be larger than would naïve 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.

Another potential weakness of *p*-curve is that, in the presence of *p*-hacking, it will underestimate a true effect. It is possible that there are detectably large effects of violent games in experiments but that the literature is contaminated by *p*-hacking, leading to downward bias in *p*-curve results.

Similarly, PET and PEESE have 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, it can be difficult to distinguish which is the more appropriate estimator. Thus, PET and PEESE might be thought of as rough estimates of lower and upper bounds on the effect, respectively, rather than 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., Anderson et al., 2004), 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 (Hilgard, Engelhardt, Bartholow, & Rouder, in press).

Application of confounds in analysis of covariance is a more promising approach, but this is also sometimes controversial (Miller & Chapman, 2001). 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 et al., 2013; Elson & Quandt, 2014; Engelhardt, Hilgard, & Bartholow, 2015; Engelhardt, Mazurek, et al., 2015; Hilgard, 2015; Kneer et al., in press), which manipulate violent content while preserving the content of gameplay (rules, controls, level design, etc.). 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 likely that there are detectable longitudinal effects of many hours of gameplay over time. 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, we have poorly-controlled manipulations yielding uncertain effects overstated through research bias.

(Mis)understanding moderators in experiments.

> "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."
> (Warburton, 2014, p. 62)

Overconfidence in the main effect leads also to overconfidence in interactions and moderators. 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. We are concerned that the understanding of nuance is overstated.

First, tests of moderators are likely to be *underpowered*. 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. If power is so poor, the positive predictive value of significant interactions is minimal; such significant interactions would be more likely to be Type I errors than true phenomena.

Second, we suspect that some tests of moderators are *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, Engelhardt, Hilgard, & Bartholow, 2015), but they should be replicated before taken as fact. One of us has published such an interaction, trait anger × violent game exposure (Engelhardt, Bartholow, & Saults, 2011), and later found that it did not replicate (Engelhardt, Hilgard, Clark, & Mazurek, n.d.). The diversity of reported moderators and the infrequency of their replication suggest possible weaknesses in the literature.

Of course, it is possible that there exist subgroups in which the effect size in experiments is larger and may be productively studied. We ask that researchers consider the aforementioned pitfalls and adopt appropriate safeguards. Pre-registration would help clarify which results are confirmatory and which are exploratory. Larger sample sizes would increase the evidentiary value of individual studies. Replication would help to identify which moderators are reliable and which are attributable to chance. 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.

Summary

In short, the research literature as analyzed by Anderson et al. (2010) 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 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.”

References

Adachi, P. J. C., & Willoughby, T. (2011). The effects of video game competition and violence on aggressive behavior: Which characteristic has the greatest influence? *Psychology of Violence*, *1*(4), 259-274. Retrieved from 10.1037/a0024908

Anderson, C. A., Carnagey, N. L., Flanagan, M., Benjamin, J., A. J., Eubanks, J., & Valentine, J. C. (2004). Violent video games: Specific effects of violent content on aggressive thoughts and behavior. *Advances in Experimental Social Psychology*, *36*, 199-249.

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., ... Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in eastern and western countries: A meta-analytic review. *Psychological Bulletin*, *136*(2), 151-173. Retrieved from http://psycnet.apa.org/doi/10.1037/a0018251

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543-554. Retrieved from 10.1177/1745691612459060

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425. Retrieved from http://dx.doi.org/10.1037/a0021524

Brady, S. S., & Matthews, K. A. (2006). Effects of media violence on health-related outcomes among young men. *Archives of Pediatric and Adolescent Medicine*, *160*, 341-347. Retrieved from 10.1001/archpedi.160.4.341

Brooks, M. C. (2000). *Press start: Exploring the effects of violent video games on boys* (Unpublished doctoral dissertation). (Dissertation Abstracts International: Section B. The Sciences and Engineering, 60(12), 6419.)

Bushman, B. J., & Huesmann, L. R. (2014). Twenty-five years of research on violence in digital games and aggression revisited: A reply to elson and ferguson (2013).

European Psychologist, *19*, 47-55. Retrieved from 10.1027/1016-9040/a000164

Bushman, B. J., Rothstein, H. R., & Anderson, C. A. (2010). Much ado about something: Violent video game effects and a school of red herring: Reply to ferguson and kilburn (2010). *Psychological Bulletin*, *136*, 182-187. Retrieved from 10.1037/a0018718

Carnagey, N. L., & Anderson, C. A. (2005). The effects of reward and punishment in violent video games on aggressive affect, cognition, and behavior. *Psychological Science*, 882-889.

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

Cohn, L. B. (1995). *Violent video games: Aggression, arousal, and desensitization in young adolescent boys* (Unpublished doctoral dissertation). University of Southern California, Los Angeles.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455-463. Retrieved from 10.1111/j.0006-341X.2000.00455.x

Egger, M. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*, 629-634. Retrieved from DOI: 10.1136/bmj.315.7109.629

Elson, M., Bruer, J., Van Looy, J., Kneer, J., & Quandt, T. (2013). Comparing apples and oranges? evidence for pace of action as a confound in research on digital games and aggression. *Psychology of Popular Media Culture*, No pagination specified. Retrieved from http://dx.doi.org/10.1037/ppm0000010

Elson, M., & Ferguson, C. J. (2013). Twenty-five years of research on violence in digital games and aggression: Empirical evidence, perspectives, and a debate gone astray. *European Psychologist*, *19*. Retrieved from DOI: 10.1027/1016-9040/a000147

Elson, M., Mohseni, M. R., Breuer, J., Scharkow, M., & Quandt, T. (2014). Press crtt to measure aggressive behavior: The unstandardized use of the competitive reaction

time task in aggression research. *Psychological Assessment*, *26*(2), 419-432. Retrieved from 10.1037/a0035569

Elson, M., & Quandt, T. (2014). Digital games in laboratory experiments: Controlling a complex stimulus through modding. *Psychology of Popular Media Culture*. Retrieved from http://psycnet.apa.org/doi/10.1037/ppm0000033

Engelhardt, C. R., Bartholow, B. D., & Saults, J. S. (2011). Violent and nonviolent video games differentially affect physical aggression for individuals high vs. low in dispositional anger. *Aggressive Behavior*, *37*, 539-546. Retrieved from 10.1002/ab.20411

Engelhardt, C. R., Hilgard, J., & Bartholow, B. D. (2015). Acute exposure to difficult (but not violent) video games dysregulates cognitive control. *Computers in Human Behavior*, *45*, 85-92. Retrieved from 10.1016/j.chb.2014.11.089

Engelhardt, C. R., Hilgard, J., Clark, K. E., & Mazurek, M. O. (n.d.). *The general aggression model: A model that does not explain the effects of violent video game exposure on aggression in adults with and without autism spectrum disorder.* (In preparation)

Engelhardt, C. R., Mazurek, M. O., Hilgard, J., Rouder, J. N., & Bartholow, B. D. (2015). Effects of violent-video-game exposure on aggressive behavior, aggressive-thought accessibility, and aggressive affect among adults with and without autism spectrum disorder. *Psychological Science*. Retrieved from 10.1177/0956797615583038

Ferguson, C. J. (2007). Evidence for publication bias in video game violence effects literature: A meta-analytic review. *Aggression and Violent Behavior*, *12*, 470-482. Retrieved from 10.1016/j.avb.2007.01.001

Ferguson, C. J., & Kilburn, J. (2009). The public health risks of media violence: A meta-analytic review. *The Journal of Pediatrics*, *154*(5), 759-763. Retrieved from 10.1016/j.jpeds.2008.11.033

Graybill, D., Kirsch, J. R., & Esselman, E. D. (1985). Effects of playing violent versus

nonviolent video games on the aggressive ideation of aggressive and nonaggressive children. *Child Study Journal*, *15*, 199-205.

Higgins, J. P. T., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions* (Vol. Version 5.1.0 [updated March 2011]). The Cochrane Collaboration. Retrieved from www.cochrane-handbook.org

Hilgard, J. (2015). *Game violence, game difficulty, and 2d:4d digit ratio as predictors of aggressive behavior* (Doctoral dissertation, University of Missouri). Retrieved from https://osf.io/uw3z8/

Hilgard, J., Engelhardt, C. R., Bartholow, B. D., & Rouder, J. N. (in press). How much evidence is p > .05? stimulus pre-testing and null primary outcomes in violent video games research. *Psychology of Popular Media Culture*. Retrieved from https://github.com/Joe-Hilgard/VG2-Bayes/

Hoffman, K. D. (1994). *Effects of playing versus witnessing video game violence on attitudes toward aggression and acceptance of violence as a means of conflict resolution* (Unpublished doctoral dissertation). University of Alabama, Tuscaloosa.

Huesmann, L. R. (2010). Nailing the coffin shut on doubts that violent video games stimulate aggression: Comment on anderson et al. (2010). *Psychological Bulletin*, *136*, 179-181. Retrieved from 10.1037/a0018567

Kneer, J., Elson, M., & Knapp, F. (in press). Fight fire with rainbows: The effects of displayed violence, difficulty, and performance in digital games on affect, aggression, and physiological arousal. *Computers in Human Behavior*. Retrieved from 10.1016/j.chb.2015.07.034

Konijn, E. A., Nije Bijvank, M., & Bushman, B. J. (2007). I wish i were a warrior: The role of wishful identification in the effects of violent video games on aggression in adolescent boys. *Developmental Psychology*, *43*(4), 1038-1044. Retrieved from http://psycnet.apa.org/doi/10.1037/0012-1649.43.4.1038

Lakens, D., Hilgard, J., & Staaks, J. (in press). On the reproducibility of meta-analyses:

Six practical recommendations. *BioMed Central*. Retrieved from http://tinyurl.com/LakensHilgardStaaks

Matsuzaki, N., Watanabe, H., & Satou, K. (2004). Educational psychology of the aggressiveness in the video game. *Bulletin of the Faculty of Education, Ehime University*, *51*(1), 45-52.

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, *110*(1), 40-48. Retrieved from http://psycnet.apa.org/doi/10.1037/0021-843X.110.1.40

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Panee, C. D., & Ballard, M. E. (2002). High versus low aggressive priming during video-game training: Effects on violent action during game play, hostility, heart rate, and blood pressure. *Journal of Applied Social Psychology*, *32*(12), 2458-2474. Retrieved from DOI: 10.1111/j.1559-1816.2002.tb02751.x

Przybylski, A. K., Deci, E. L., Rigby, C. S., & Ryan, R. M. (2014). Competence-impeding electronic games and players' aggressive feelings, thoughts, and behaviors. *Journal of Personality and Social Psychology*, *106*(3), 441-457. Retrieved from http://psycnet.apa.org/doi/10.1037/a0034820

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple–study articles. *Psychological Methods*, *17*, 551–566.

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

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534-547. Retrieved from 10.1037/a0033242

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). P-curve and effect size:

Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*, 666-681. Retrieved from 10.1177/1745691614553988

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60-78. Retrieved from DOI: 10.1002/jrsm.1095

Viechtbauer, W. (n.d.). Conducting meta-analyses in R with the metafor package.

Warburton, W. (2014). Apples, oranges, and the burden of proof: Putting media violence findings into context: A comment on elson and ferguson (2013). *European Psychologist*, *19*, 60-67. Retrieved from 10.1027/1016-9040/a000166

Willoughby, T., Adachi, P. J. C., & Good, M. (2012). A longitudinal study of the association between violent video game play and aggression among adolescents. *Developmental Psychology*, *48*, 1044-1057. Retrieved from 10.1037/a0026046

Table 1

PET, PEESE, and p-curve adjusted estimates.

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

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**

Table 3

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

Liberal coding scheme.

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

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

Conservative coding scheme.

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

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

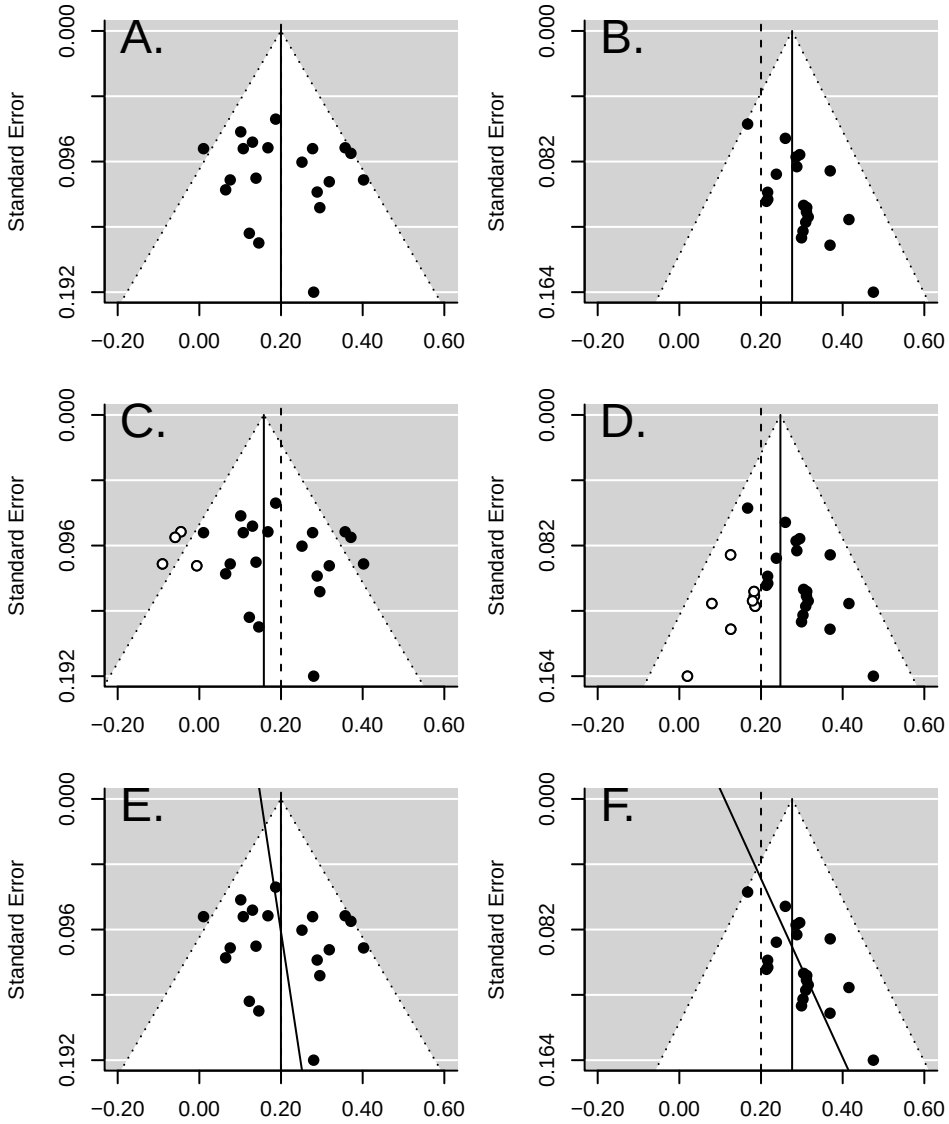

Figure 1. Funnel plots, trim-and-fill, and Egger's test. Effect size Fisher's z is on the x-axis, while standard error of Fisher's z is on the y-axis. The true effect size $z = .2$ is indicated by the dashed line. Panels A and B show funnel plots for unbiased and biased literatures, respectively. The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

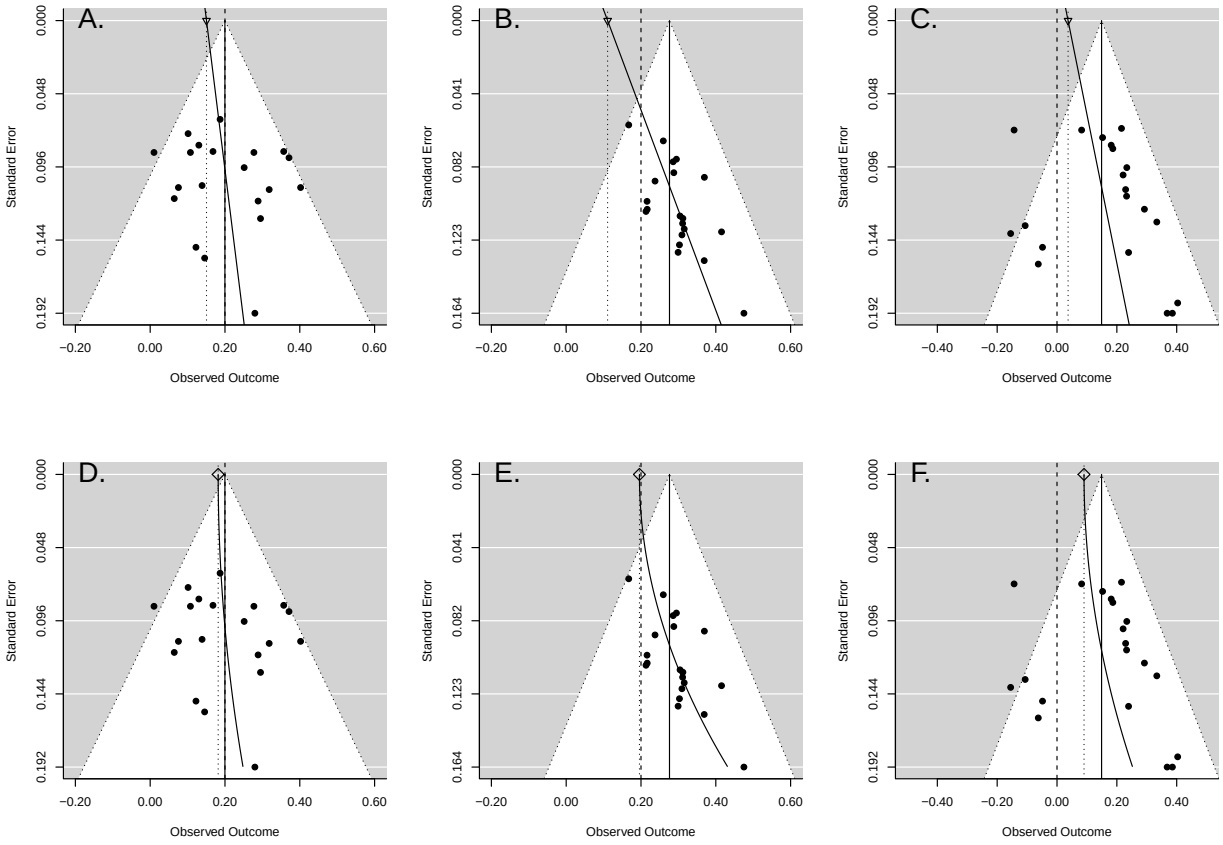

Figure 2. PET and PEESE meta-regression. Again, Fisher's z is on the x-axis, standard error is on the y-axis, and the true effect size is indicated by the dashed line. Bias-adjusted estimates are indicated by the dotted vertical line. Panels A and B indicate the PET technique applied to unbiased and 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.

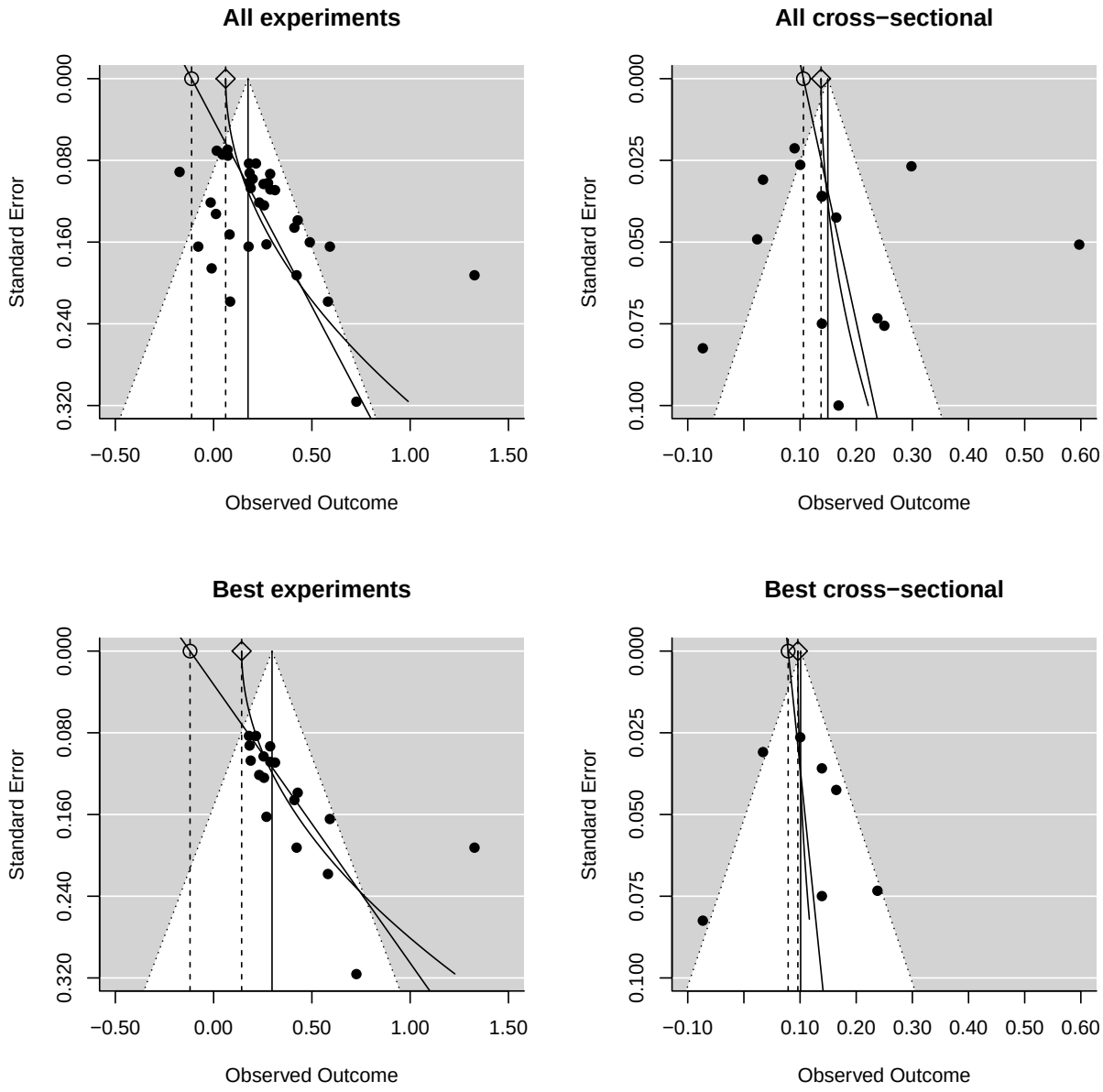

Figure 3. Funnel plot of studies of aggressive affect with overlaid PET and PEESE meta-regression lines.

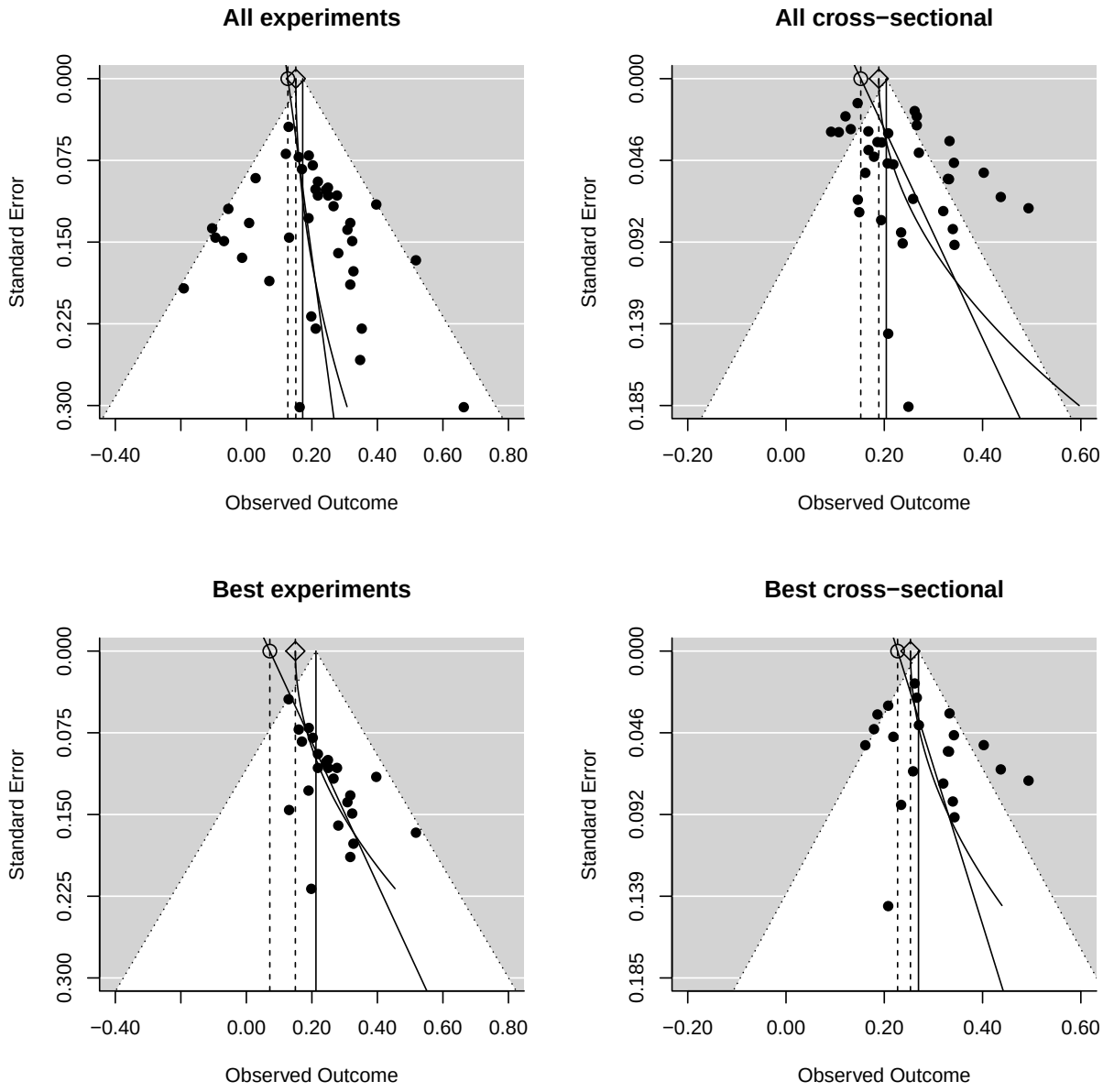

Figure 4. Funnel plot of studies of aggressive behavior with overlaid PET and PEESE meta-regression lines.

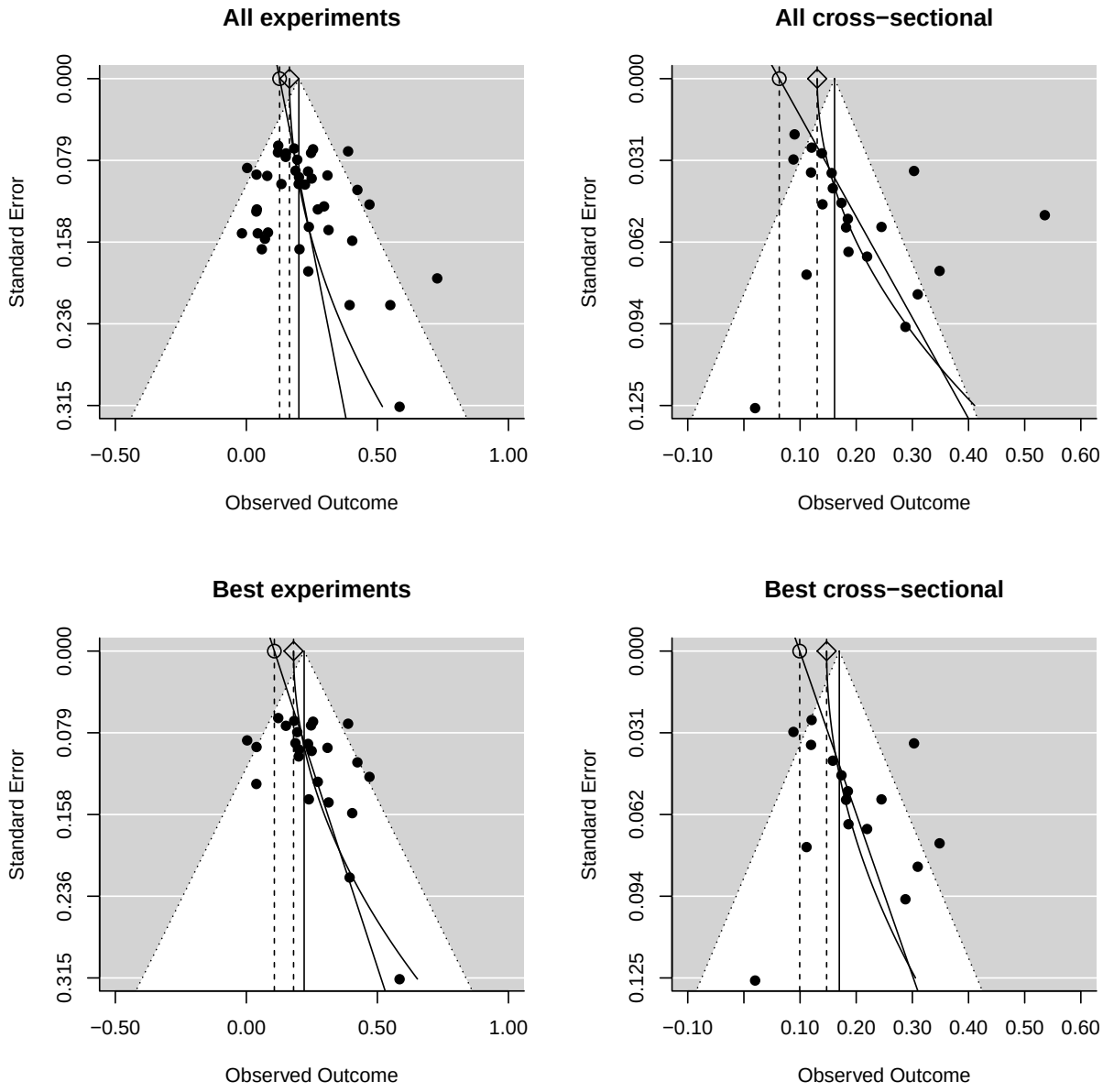

Figure 5. Funnel plot of studies of aggressive cognition with overlaid PET and PEESE meta-regression lines.

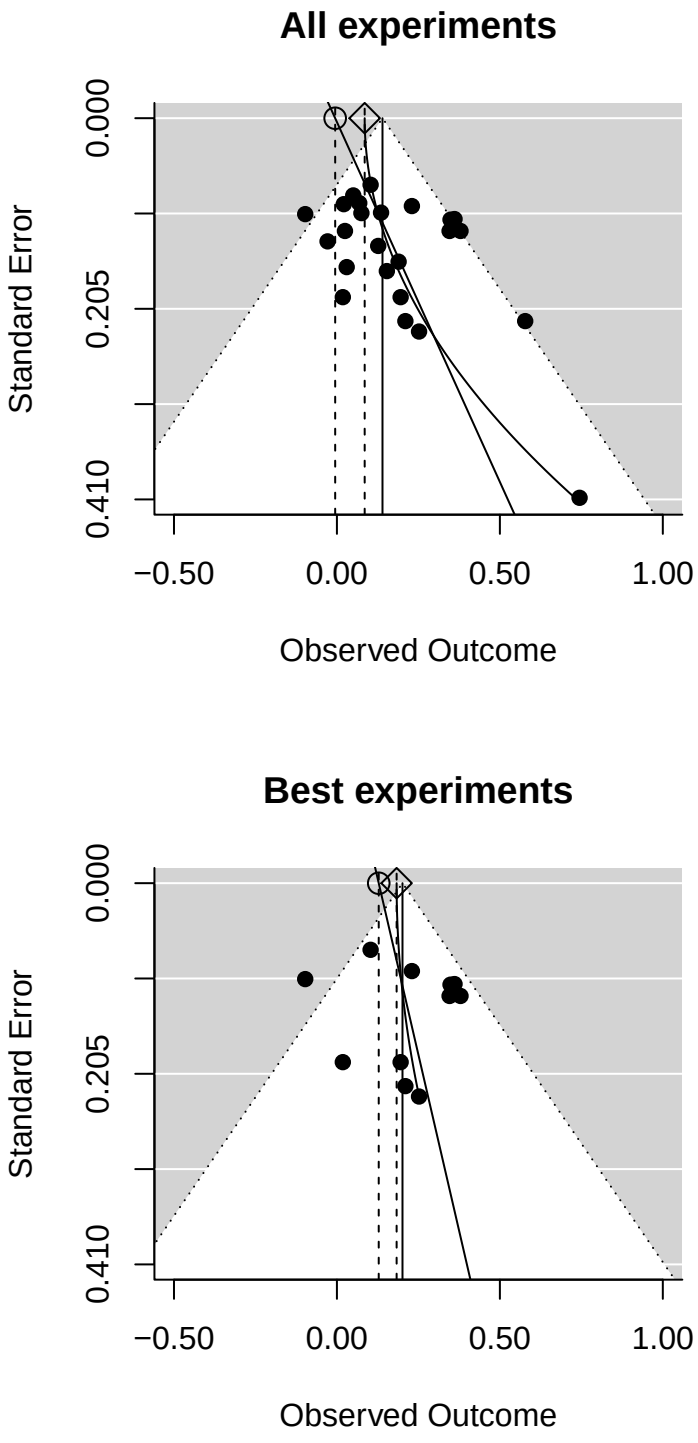

Figure 6. Funnel plot of studies of physiological arousal with overlaid PET and PEESE meta-regression lines.

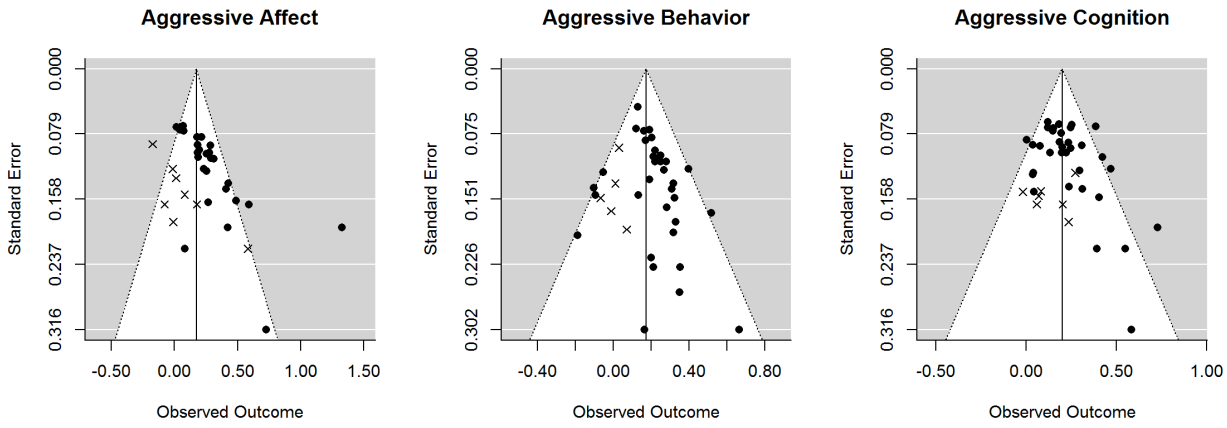

Figure 7. Funnel plots of all experiments of aggressive affect, behavior, and cognition. Dissertations not presented in any further publication format are indicated with Xs, while all other publication styles (e.g., journal articles, book chapters, conference proceedings) are indicated with filled dots.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