Overstated Evidence for Short-Term Effects of Violent Games on Affect and Behavior: A Reanalysis of Anderson et al. (2010)

Joseph Hilgard
University of Pennsylvania

Christopher R. Engelhardt
CARFAX, Inc.

Jeffrey N. Rouder
University of Missouri

Author Note

Please direct correspondence regarding this article to Joseph Hilgard. E-mail: jhilgard@gmail.com

We thank Craig A. Anderson for sharing with us the dataset from Anderson et al. (2010) and inviting us to host it publicly in our GitHub repository. We thank Randy McCarthy and Katie Corker for suggestions on an earlier draft of this manuscript. Joseph Hilgard is supported by the Drs. Gloria and Melvin "Jack" Chisum Research Fellowship at the Annenberg Public Policy Center. Jeffrey N. Rouder is supported by National Science Foundation Grants BCS-1240359 and SES-102408.

THIS MANUSCRIPT HAS NOT BEEN PEER-REVIEWED. DO NOT CITE WITHOUT THE PERMISSION OF THE CORRESPONDING AUTHOR.

Abstract

Violent video games are theorized to be a significant cause of aggressive thoughts, feelings, and behaviors. Important evidence for this claim comes from a large meta-analysis by Anderson and colleagues (2010), who found effects of violent games 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. In the present manuscript, we re-examine their meta-analysis using a wider array of techniques for detecting bias and adjusting effect sizes. Our conclusions differ from those of Anderson and colleagues in three salient ways. First, we detect substantial publication bias in experimental research on the effects of violent games on aggressive affect and aggressive behavior. Second, after adjustment for bias, the effects of violent games on aggressive behavior in experimental research are estimated as being very small, and estimates of effects on aggressive affect are much reduced. In contrast, the cross-sectional literature finds correlations that appear largely unbiased. Third, experiments meeting the original authors' criteria for methodological quality do not yield larger adjusted effects than other experiments, but instead yield larger indications of bias, indicating that perhaps they were selected for significance. 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 technological sophistication, even modest effects of violent games could have serious implications for public health. Psychological research provides seemingly strong evidence of such a link, so much so that professional organizations have issued policy statements describing 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. As the American Academy of Pediatrics puts it, "the association between media violence and aggressive behavior is […] nearly as strong as the association between cigarette smoking and lung cancer" (AAP, 2009, p. 1497).

Despite decades of research and hundreds of studies, however, the basic phenomena remain debated. 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 (see, for example, Elson & Ferguson, 2014).

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 covered studies 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. They find 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 are not limited to experiments but are also found in cross-sectional comparisons and even in longitudinal research designs. Of course, the quality of any meta-analysis may be compromised by publication bias. To mitigate the effects of such bias, Anderson et al. applied the trim-and-fill procedure (Duval & Tweedie, 2000) to detect and adjust for the effects of publication bias on effect size estimates. According to Anderson et al., the resulting adjustments were minimal, indicating there was no serious bias among the studies. With these safeguards in place, Anderson and colleagues conclude there is strong evidence for violent-video-game effects; 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. Ferguson and Kilburn (2010), for example, are concerned that the Anderson et al. (2010) meta-analysis may suffer from biases in the publication of studies, the entry of effect sizes into meta-analysis, and the application of the best-practices inclusion criteria. Other skeptics, such as Elson, Mohseni, Breuer, Scharkow, and Quandt (2014), are also concerned that the individual studies suffer from questionable research practices such as the selective report of dependent variables that yield statistical significance. Skeptics suspect that these meta-analytic biases and questionable research practices may overestimate the strength of evidence for, and magnitude of, violent video game effects, despite the results of trim-and-fill analyses.

To address this continued skepticism, we re-analyze the database created and used by Anderson et al. (2010). 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 assessment of the violent-video game literature has important theoretical ramifications. The leading explanation of these effects, Anderson and Bushman's General Aggression Model (GAM, 2002) is a conventional social-psychological activation model in which violent games increase arousal, cause aggressive affect, and prime aggressive thoughts, which, in turn, cause aggressive behavior. Violent video game effects therefore are anticipated by popular theoretical mechanisms in social psychology. Third, the Anderson et al. (2010) meta-analysis is a tremendous volume of work encompassing many studies. We were drawn to the quality and quantity of data, as well as its considerable impact on theory, practice, and policy. Fourth, there are promising new techniques for addressing potential publication bias and questionable research practices. These new techniques include PET (Precision-Effect Test; Stanley & Doucouliagos, 2014), PEESE (Precision-Effect Estimate with Standard Error; Stanley & Doucouliagos, 2014), *p*-curve (Simonsohn, Nelson, & Simmons, 2014a, 2014b), and *p*-uniform (van Assen, van Aert, & Wicherts, 2015). The articles introducing these techniques each perform simulations demonstrating better adjustments for these potential artifacts than the trim-and-fill method used in Anderson et al. (2010). Application of these techniques, then, may yield new insights regarding the magnitude of effects on certain outcomes in certain paradigms.

Concerns about Bias

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

(e.g., $p < .05$) findings are more likely to be submitted and accepted for publication than are studies with non-significant results. 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. When studies that attain statistical significance are more likely to be published than those that are not, meta-analyses of the published literature are no longer representative of the full body of research. Note that publication bias is proportionate, not absolute. The presence of some published null results therefore does not rule out the possibility of any publication bias. Note also that the bias can be inflicted at both the level of journals, which may reject null results, and authors, who may not bother submitting null results. Meta-analyses of literatures suffering from publication bias are likely to overestimate effect sizes and may reach incorrect conclusions of statistically and practically significant effects.

The critical question is whether there is evidence for publication bias in the violent video-game literature as synthesized by Anderson et al. (2010). Here there is disagreement. Anderson et al. claim that there is little evidence for publication bias. Their claim follows from their attempts to account for such bias using both statistical methods and literature review.

With regard to statistical methods, the authors used a trim-and-fill procedure 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, although trim-and-fill was quite popular at the time of the Anderson et al. analysis, today we

understand trim-and-fill to be at least somewhat flawed. It corrects for bias when bias is absent and does not correct enough when bias is strong (Simonsohn et al., 2014b; van Assen et al., 2015). It also has difficulty adjusting effect sizes to zero when the null is true and there is publication bias (Moreno et al., 2009; van Assen et al., 2015).

With regard to literature review, the authors made an attempt to collect unpublished literature. The authors found 18 dissertations that had gone unpublished, 16 of which failed to find statistical significance on one or more outcomes. Only one unpublished non-dissertation study was found. Given the difficulty of gathering unpublished results, we suspect that there may be more unpublished non-dissertation studies censored from report. 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. Practices that lead to this inflation of Type I error 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 analysis of the full sample fails to detect an effect. Another form of *p*-hacking is outcome switching: If an experiment's primary outcome does not find the desired result, other outcomes with more statistically-significant changes might be presented instead and the primary outcome hidden from report.

It has been argued that such outcome-switching may exist in the quantification and report of certain measures of aggressive behavior. Some researchers measure aggressive behavior by allowing participants to administer a painful burst of noise to another participant (the

Competitive Reaction Time Task). 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 Elson et al. (2014) suggest that this 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. Such selection of the larger, more statistically significant outcomes risks overestimation of the net effect size.

Selection bias. Selection bias may contaminate meta-analysis when the researchers include or exclude studies on the basis of the hypothesis they favor. To their credit, Anderson et al. define a set of "best practices" inclusion criteria. Studies meeting these criteria are argued to provide better estimates through higher methodological quality, say, by proper operationalization of the independent variable, or by an appropriate choice of outcome measure (Anderson et al., 2010, Table 2). However, Ferguson and Kilburn (2010) took issue with Anderson et al.'s best practices criteria. These critics argued that the inclusion criteria were applied more liberally to studies with significant results than to studies with nonsignificant results, thereby artificially inflating the evidence for, and size of, the effect. Given the controversy, we will pay close attention to how results vary with and without these best-practices criteria.

Assessing Bias in Meta-Analysis

There are several approaches to assessing the aforementioned biases in meta-analysis. Some of these are recent developments published only after the publication of Anderson et al. (2010). 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 sizes and standard errors (or sample size) in reported studies. In an unbiased research literature, there should be no relationship between effect sizes and standard errors; sample size does not cause effect size. However, such a relationship will be observed if publication favors statistically-significant studies at the expense of nonsignificant studies. Small-sample studies need large observed effect sizes to reach statistical significance, whereas large-sample studies can reach statistical significance with smaller observed effect sizes. Thus, in the presence of publication bias, there is a *small-study effect:* an inverse relationship between effect size and sample size.

We apply a few modern methods for detecting small-study effects that, as we will document, indicate substantial bias in the Anderson et al. meta-analysis. One critical issue in meta-analysis is the question of what may be learned in the presence of bias. The most charitable position is that researchers may assess the degree of bias and provide needed corrections to recover accurate effect size estimates (e.g., Duval & Tweedie, 2000; Simonsohn et al., 2014b). We are less sanguine, as much is unknown about the statistical properties of corrections—their efficiency and bias in realistically-sized samples as well as their robustness to violations of assumptions. Still, they have some value in analysis. We provide a review of the bias-detection-and-correction methods used in this study, noting the strengths and weaknesses of each.

Funnel plots. Funnel plots provide a useful graphical summary of potential small-study effects in meta-analysis. The relationship between effect sizes and standard errors is plotted, allowing for visual estimation of small-study effects. In a funnel plot, effect size is plotted on the *x*-axis and precision (that is, the inverse of the standard error) on the *y*-axis. In the absence of small-study effects and 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 no bias.

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 fail to reach statistical significance or would even suggest an effect of opposite sign. This asymmetry can also be caused by *p*-hacking. 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 censored 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 highlight only the significant outcomes, the effect size increases, moving studies to the right of the funnel.

Egger's regression test. Egger's weighted regression test (Egger, Smith, Schneider, & Minder, 1997) inspects the degree and statistical significance of the relationship between standard errors and effect sizes. 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.

Egger's regression test has some weaknesses. Although 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, & Egger, 2000). Performance is also likely to

degrade under conditions of heterogeneity (e.g., Lau et al., 2006; Terrin et al., 2003). Skeptics have used Egger's test to look for evidence of bias in the violent-game-effect literature (e.g., Ferguson, 2007; Ferguson & Kilburn, 2009), but Anderson et al. (2010) abstained from its use.

Trim and fill. One popular bias-adjustment technique, trim and fill (Duval & Tweedie, 2000), is used 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 of the absolute values of the observed effect sizes on each side of the mean effect size are roughly equal.

Trim-and-fill has its critics. Moreno et al. (2009), Simonsohn et al. (2014b), and van Assen et al. (2015) argue it is not useful: when there is no bias, there is too much adjustment, and when there is strong bias, there is too little adjustment. Higgins and Green (2011) express concern about the imputation of studies, which adds purely hypothetical data to the meta-analysis.

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 the trim-and-fill procedure in their meta-analysis. The procedure yielded only slightly-adjusted effect sizes, and so the authors concluded minimal research bias.

Again, the development of novel adjustments for small-study effects allows for further testing of this conclusion.

PET and PEESE meta-regression. Meta-regression is a promising new tool in bias detection and adjustment. Meta-regression estimates a bias-adjusted effect size by considering the relationship between effect sizes and standard errors, then estimating the hypothetical underlying effect size that would be found if the standard error were zero. 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 sizes and standard errors, as in the Egger regression test. Unlike Egger's test, which considers the slope of this regression, PET considers the intercept of this regression. This extrapolates from the available data to estimate what the effect would be in a hypothetical study with perfect precision. When there is minimal bias, there is minimal adjustment. 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. However, PET performs less well when there is some effect. When there is an underlying effect, small studies will be censored heavily 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 (Stanley & Doucouliagos, 2014).

A second meta-regression estimator, PEESE, is intended to address this problem. PEESE fits a weighted *quadratic* relationship between effect sizes and standard errors. The motivation for the additional quadratic term is as follows: Assuming there is some true effect, and that

publication bias favors statistically-significant results, poorly-powered, low-precision studies will be publishable only when they badly overestimate the true effect size. In contrast, well-powered, high-precision studies will routinely get significant results and will therefore be publishable without overestimating the true effect size. Thus, there is more bias among low-precision studies than there is among high-precision studies; the quadratic term allows for this distinction. In practice, PEESE is less likely than PET to underestimate nonzero effects, but risks overestimating the size of null effects (Stanley & Doucouliagos, 2014).

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 (Gervais, 2015). 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 might 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.

One recent example of the profitable use of these meta-regression techniques is a meta-analysis by Carter and McCullough (2014). These authors tested the 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, which found no evidence of ego depletion (Hagger et al., in press).

P-Curve. Another novel technique for accounting for small-study effects is *p*-curve (Simonsohn et al., 2014a, 2014b), which 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 less than .01 as they are between .04 and .05. When the null hypothesis is false, the *p*-curve becomes right-skewed such that *p*-values less than .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 mean population effect size.

P-curve also has the weakness is that can return biased estimates when individual studies are *p*-hacked. Simonsohn et al. (2014b) warn that *p*-hacking is likely to cause *p*-curve to underestimate the effect size, but van Aert et al. (in press) warn that, under certain conditions, *p*-hacking can cause *p*-curve to overestimate the mean population effect size.

P-uniform. *P*-uniform is another power-based test and adjustment for bias (van Assen et al., 2015). Like *p*-curve, it considers only the statistically-significant results in meta-analysis. It attempts to find an underlying effect size for which the conditional *p*-value distribution would be as close to uniform as possible. That is, it looks for an effect size δ_0 for which the null hypothesis H_0: $\delta = \delta_0$ would generate an approximately uniform distribution of *p*-values. It also provides a test for publication bias by considering whether the adjusted effect size is statistically

significantly smaller than the naïve meta-analytic estimate. Like *p*-curve, it only considers studies with $p < .05$, and so may lose substantial information.

Unpublished Dissertations. Yet another approach is to eschew statistical adjustments and attempt to inspect the unpublished literature directly. When unpublished work provides smaller effect size estimates than published work, publication bias may be present. Unfortunately, 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 (O'Boyle, Banks, & Gonzalez-Mule, 2014). Dissertations, then, provide us with a sample of reported studies relatively uncontaminated by publication biases favoring significant results. In our analyses, we examine these unpublished dissertations and the statistical significance of their results.

Heterogeneity

One critical issue in meta-analysis is whether the gathered effect sizes from individual studies seem to represent a single underlying effect size. Often, effect sizes vary by more than sampling error alone would predict, indicating that the effect is sometimes larger or smaller from study to study. This variance is called heterogeneity. Sometimes this heterogeneity indicates that studies are too dissimilar to be combined and should instead be considered separately; for example, cross-sectional and experimental research may reflect different phenomena and should

not be combined. At other times, heterogeneity in effect sizes can be predicted as a function of some moderator through meta-regression. For example, an effect might be larger among males than among females. Such meta-regression findings are helpful in that they provide new discoveries and reduce the amount of residual heterogeneity. Finally, sometimes heterogeneity is present and cannot be attributed to any moderator. In this unfortunate but common situation (Higgins, 2008), meta-analytic conclusions should be interpreted with the understanding that there are as-yet unknown factors that cause the population effect size to be larger or smaller from study to study.

Sometimes heterogeneity can cause a correlation between effect sizes and standard errors that is not due to bias. For example, experimental studies tend to have smaller samples than cross-sectional studies, and each paradigm may reflect different underlying effect sizes. In this case, sample size would be related to effect size, but only through the confounding variable of study design. It is desirable to rule out such potential confounds so that small-study effects more likely represent the role of bias. For example, conducting separate bias tests for cross-sectional and experimental studies can rule out study design as a potential cause of small-study effects.

Heterogeneity is of particular concern in the present report, as most of the bias-adjustment methods we apply assume homogeneity, and all will have difficulty in the face of substantial heterogeneity. Under conditions of heterogeneity, funnel plots may overestimate the degree of asymmetry (Lau, Ioannidis, Terrin, Schmid, & Olkin, 2006; Terrin, Schmid, Lau, & Olkin, 2003). Variability among studies may cause some precisely estimated studies to have effect size estimates far from the overall mean, leading to Type I or Type II errors in bias tests. Thus, performance of funnel-plot-based tests and adjustments such as the Egger test, trim-and-

fill procedure, and PET and PEESE meta-regression will degrade in the presence of heterogeneity.

Heterogeneity can also influence power-based (i.e., *p*-curve, *p*-uniform) meta-analytic adjustments. These techniques consider only the statistically significant results, and studies with greater underlying effect sizes are more likely to attain statistical significance. This can cause *p*-curve and *p*-uniform to estimate a larger effect size than does naïve meta-analysis, as the naïve analysis considers all studies, whereas *p*-curve and *p*-uniform consider only the statistically-significant studies. Van Aert, Wicherts, and van Assen (in press) caution that this will lead to overestimation of effect sizes under moderate to large heterogeneity.

Anderson et al. noticed heterogeneity in several subsets in their meta-analysis. They attempted to resolve this heterogeneity by looking for moderators, and found that none of their hypothesized candidates, such as participant age or Eastern vs. Western culture, accounted for this heterogeneity. In our analyses, we look to see how much heterogeneity can be resolved as small-study effects, and caution the reader when adjusted effect sizes may be influenced by residual heterogeneity. To foreshadow, although residual heterogeneity was detected for a few subsets, treating standard errors as a covariate greatly reduced residual heterogeneity, and in one case, the resulting residuals were so homogeneous as to suggest severe contamination by bias (see Ioannidis, Trikalinos, & Zintzaras, 2006).

Summary of Methods

Given this state of the field, our analysis will consist of two main questions. First, is there evidence of small-study effects in the dataset? The presence or absence of these effects will be assessed informally by inspection of funnel plots and more formally by the Egger test and *p*-uniform's bias test. Second, what might be appropriate bias-adjusted estimates? We will apply

PET, PEESE, *p*-curve, and *p*-uniform 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. We will consider whether there are differences between the results of published articles and unpublished dissertations that might suggest bias. Finally, we will consider the degree of heterogeneity of studies as the above methods may in some cases be degraded by heterogeneity.

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, *p*-curve, and *p*-uniform analyses. We use the original authors' separation of studies by study design (experimental, cross-sectional, longitudinal), by study outcome (affect, behavior, cognition, arousal), and by study quality (all studies, best-practices subset) in our presentation. Thus, point-biserial correlations from experiments and product-moment correlations from cross-sections are treated separately, as is generally preferred. Finally, we perform χ^2 tests to see whether unpublished dissertations are more or less likely to yield statistical significance than other published work.

In the original dataset, Anderson et al. (2010) coded all effect sizes in terms of Pearson *r*, then converted these to Fisher's *z*-scores with standard error equal to $\frac{1}{\sqrt{N-3}}$.[1] This approach is appropriate given that most outcome measures are either continuous or at least modeled as continuous by study authors. We use their estimated *z*-scores and standard errors in this analysis.

[1] As a reviewer points out, this approximation is technically only correct when the effect size is zero; as the effect size increases, the standard error becomes smaller than $\frac{1}{\sqrt{N-3}}$. Still, we prefer this estimator because it eliminates an inherent correlation between effect size and standard error, thereby avoiding potential bias in meta-regression tests. Additionally, the approximation is good when effects are not too large, as here. See, e.g., Borenstein, 2009, p. 234.

This approach has the benefit of providing standard errors that are not a function of effect size. Standard errors that are a function of their corresponding effect sizes can lead to the spurious detection of small-study effects.

Our inspection focuses on the raw effect sizes contained in that report. Anderson and colleagues report partial correlation coefficients from cross-sectional studies which are adjusted for participants' sex (k = 9, 36, and 21 for affect, behavior, and cognition, respectively). We abstain from analysis of these, as re-analysis of the partial effect sizes is likely to be challenging due to the particularities of partial correlations (see, e.g., Aloe, 2014).

All data and code have been made available online at https://github.com/Joe-Hilgard/Anderson-meta.

Aggregation within Studies

As we apply them, the meta-analytic procedures assume that entire studies are censored or re-analyzed per their statistical significance. However, the original data have some studies divided into subsets to test for moderators. For example, one study might be entered as two records: 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 effect size estimate 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 the equation $r \times \sqrt{\frac{n-2}{1-r^2}}$, 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. 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 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. Egger tests, PET, and PEESE were performed using the `metafor` package for **R** (Viechtbauer, 2010), using the `rma()` function to fit a weighted random-effects model with an additive error term.[2] Models were fitted via restricted maximum-likelihood (REML) estimation, per package defaults. Effect sizes are converted from Fisher's *z* to Pearson *r* for tables and discussion.

For *p*-curve, we used the R code behind version 3.0 of the online *p*-curve app (Simonsohn et al., 2014a), entering a *t*-value and degrees of freedom parameter for each relevant study. This code provides estimates in terms of Cohen's *d*. We converted these to Pearson *r* for consistency of presentation, using the formula $r = \frac{d}{\sqrt{d^2+4}}$. Full *p*-curve output from the online p-curve.com application is available in the supplementary materials

For *p*-uniform, we use the `puniform` package provided by van Aert at https://github.com/RobbievanAert/puniform. Analysis was performed using the correlations and sample sizes as entered by Anderson et al. The package's default method for the aggregation of *p*-values was used.

[2] We also fit fixed-effects models with a multiplicative error term. See the supplement.

PET, PEESE, and *p*-curve are likely to perform poorly when there are few datapoints. Therefore, our analyses are restricted to effects and experimental paradigms with at least ten independent effect sizes. Readers wanting to generate estimates for more sparse datasets or explore the impact of our inclusion and exclusion decisions are invited to download the data and code.

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.

Studies Excluded

We removed two studies from the meta-analytic database due to concerns over relevance. Panee and Ballard (2002) was removed because the study tested the effects of a violent or nonviolent training level on two outcomes: self-reported affect and aggressive behaviors within a violent video game. All participants played the same violent game; therefore, it does not provide a relevant test of the hypothesis. Graybill, Kirsch, and Esselman (1985) was also removed from analysis, as this study measured not the amount of aggressive cognition, but the direction and type of it. Because each subject was categorized into one directional and one typological category, the results do not estimate differences in the quantity of aggressive cognition. As entered in the Anderson et al. dataset, the study's manipulation checks were also entered as though they were primary study outcomes on aggressive cognitions. Neither of these are hypothesis-relevant tests.[3]

[3] In their original report, Anderson et al. (2010) report trim-and-fill analyses only for the "best practices" experiments and "best partials" cross-sections. Of these exclusions, only Panee and Ballard (2002) has any effect sizes entered as best-practices experiments (one, aggressive affect). We tested the degree to which this exclusion changed the results of naïve and trim-and-fill analysis. Even without this exclusion we were unable to reproduce

Subsets Re-analyzed

We reproduce estimates from Anderson et al. (2010) and apply PET, PEESE, *p*-curve, and *p*-uniform 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. As much as sample sizes permitted, studies were further divided to create separate best-practices-only and all-studies estimates per Anderson et al. (2010) as sample sizes permit.

The numbers of studies, overall numbers of participants, and naïve fixed- and random-effects estimates are provided for each subset in Table 1.

Results

The data set for analysis was comprised of the following subsets: I. Experimental effects of violent-game exposure on aggressive affect. In one analysis, all studies were included (k = 34, N = 2879), in another only Anderson et al.'s best-practices studies were included (k = 18, N = 1318); II. Experimental effect of violent-game exposure on aggressive behavior (all studies, k = 39, N = 3328; best practices, k = 23, N = 2413), III. Experimental effects of violent-game exposure on aggressive cognitions (all studies, k = 40, N = 4074; best practices, k = 24, N = 2887), and IV. Experimental effect of violent-game exposure on physiological arousal (all studies, k = 24, N = 1770; best practices, k = 11, N = 833). Additionally, there were enough studies to re-analyze the correlations between violent game play and aggressive affect (all studies, k = 14, N = 9811; best practices, k = 7, N = 4348, too few to re-analyze), behavior (all

their trim-and-fill result for aggressive affect: they report $r+$ = .294, with zero imputed studies, whereas we get $r+$ = .247, with six studies imputed to the left side of the funnel plot. See the supplement for details.

studies, $k = 37$, $N = 29113$; best practices, $k = 22$, $N = 12391$), and cognitions (all studies, $k = 22$, best practices $N = 13012$; best, $k = 17$, $N = 7997$) in non-experimental cross-sections.

The target of our analysis is whether there are small-study effects indicative of bias, and if so, what would be appropriate bias-adjusted effect size estimates. We present each in turn.

Detection of Bias

The first question is addressed by inspection of the funnel plots in Figures 1, 2, 3, and 4. We find dramatic funnel-plot asymmetry among experiments of aggressive affect and among best-practices experiments of aggressive behavior. Among these subsets, application of best-practices criteria seems to have exaggerated, rather than ameliorated, the funnel-plot asymmetry.

This funnel-plot asymmetry was tested by Egger's regression. Results are provided in Table 2. The regression test for funnel-plot asymmetry was statistically significant in the full set and best-practices subset of experiments studying aggressive affect. Additionally, 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. *P*-uniform also suggested significant bias for experiments of aggressive behavior, both for the full sample and for the best-practices subsample.

In total, there is clear evidence of small-study effects in studies of certain violent-game effects. This result indicates that the collected meta-analytic data may be contaminated by publication, analytic, or selection biases, and may therefore yield biased overestimates of effect sizes.

Adjusted Effect Sizes

Table 3 reports bias-adjusted effect sizes. We find that the cross-sectional studies show little need of adjustment for bias. However, the experiments received substantial adjustments; we describe those adjustments below.

In experiments of aggressive affect, the original report suggested no adjustment was necessary for the best-practices subset. In contrast, our analyses suggested downward adjustments. Relative to the fixed-effects estimate, *p*-uniform suggested an adjustment of -.05 to r = .24, and *p*-curve suggested an adjustment of -.08 to r = .21. PEESE adjusted by -.15 to r = .14, and PET adjusted the effect into the opposite direction (r = -.12).[4] We do not interpret this result an indication that the effect is literally of the opposite sign, but rather we see it as an overzealous hyperadjustment by PET in the presence of severe funnel-plot asymmetry. PEESE, *p*-curve, and *p*-uniform estimates were statistically significant, whereas the PET estimate was not.

In experiments of aggressive behavior, the original report suggested an adjustment of -.03 to r = .18. In contrast, our analyses recommended larger downward adjustments ranging from -.06 to -.19, reducing r to .15 (PEESE) or as little as .02 (*p*-uniform). Methods were conflicted as to whether the estimate was statistically significant: PEESE and *p*-curve indicated statistical significance, whereas PET and *p*-uniform did not. Our analyses also contest Anderson et al.'s conclusion that studies in the best-practices subsample find larger effects than do the not-best-practices studies. PEESE, *p*-uniform, and *p*-curve suggested identical estimates for the full sample and the best-practices subsample, whereas PET suggested that the effect was larger in the full sample than in the best-practices subsample. This latter result is perhaps an artifact of the increased strength of small-study effects in the best-practices subsample.

[4] One outlier had moderate influence over these results. See the supplement for a sensitivity analysis.

In experiments of aggressive cognition, the original report suggested an adjustment of -.02 to r = .20. Our adjustments are divergent, perhaps due to the moderate heterogeneity among studies of this outcome. *P*-uniform suggested increasing the estimate by .02 to r = .24, *p*-curve suggested an adjustment of -.03 to r = .19, PEESE suggested adjusting by -.04 to r = .18, and PET suggested adjusting by -.12 to r = .10. Again, PEESE, p-curve, and p-uniform estimates were statistically significant, whereas the PET estimate was not.

Estimates of the effects on physiological arousal seemed robust to adjustments for small-study effects. Among the best-practices subset of experiments, PEESE, *p*-curve, and *p*-uniform suggested effects as large as, or larger than, the naïve estimate. PEESE, *p*-curve, and *p*-uniform estimates were statistically significant, and the PET estimate was not.

Among cross-sectional studies, our estimators suggested minimal need for adjustment. PEESE, *p*-curve, and *p*-uniform all estimated effect sizes very close to the naïve random-effects estimate. However, the considerable heterogeneity in these subsets may limit the efficacy of these adjustments and may indicate the need for further consideration of differences in study methodology and populations.

There are some instances of convergence in our presented estimates. When inspecting effects on aggressive behavior in experiments, *p*-curve, *p*-uniform, and PET estimated that the underlying effects were so small as to be possibly undetectable in typical sample sizes (r = .02–.09). Notably, these estimates are highly consistent with some recent reports (Engelhardt, Mazurek, Hilgard, Rouder, & Bartholow, 2015; Kneer, Elson, & Knapp, 2016; Przybylski, Deci, Rigby, & Ryan, 2014; Tear & Nielsen, 2014). For effects on aggressive affect in experiments, *p*-curve, *p*-uniform, and PEESE yielded similar estimates once an outlier was removed, r = .17–.20. Caution is still warranted in light of the sharp funnel-plot asymmetry. Finally, among

experiments of aggressive cognition, *p*-curve, *p*-uniform, and PEESE converged on the $r = .18$–.24 range. Here too caution is necessary due to the considerable residual heterogeneity.

In summary, our analyses suggested that certain effects in experiments had likely been overestimated, sometimes badly. Effects on aggressive behavior were estimated as being small, possibly null. Effects on aggressive affect were estimated as moderate in size, although PET suggested a null effect. By contrast, effects on aggressive cognition and physiological arousal seemed less overestimated, and correlations from cross-sectional studies seemed relatively unbiased.

We caution the reader that these adjustments should be taken in context: Recall that we do not know the small-sample properties of the adjusted estimators and so do not valorize one in particular as being likely to provide the most accurate estimate of the underlying effect. We also refrain from taking too seriously the statistical (non)significance of the estimated effect sizes. Estimation adjustment in the face of substantial bias is heavily dependent on model assumptions, and we are unsure how to assess whether these assumptions hold and how robust estimates are to violations. Of course, these issues are not unique to our analyses, as they hold broadly across the field of meta-analysis.

Residual heterogeneity. Modeling the relationship between standard errors and effect sizes also substantially reduced the residual heterogeneity in some subsets of the data. Among best-practices experiments of aggressive affect, no heterogeneity remained in the PET and PEESE models. Similar homogeneity was attained among experiments of aggressive behavior in both the best-practices and full samples. This suggests that there is little residual variance in study results that could productively be explained by study attributes. In the case of best-practices experiments of aggressive behavior, there was so little residual variance that a

confidence interval on I^2 consisted of the null/empty set. The documentation for `metafor` suggests that this indicates "highly (or overly) homogeneous data," (Viechtbauer, 2010, helpfile for `confint.rma.uni`) an unusual absence of residual sampling variance. This would be consistent with the presence of bias: Effect sizes in this subset seem to reach statistical significance with improbably high precision (Ioannidis et al., 2006).

By comparison, modest heterogeneity remained among experiments of aggressive cognition and among the full sample of experiments of aggressive affect. Heterogeneity was also present among nonexperimental work, particularly in studies of aggressive affect. More work will be necessary to determine what distinguishes those studies finding larger effects from those finding smaller effects.

Unpublished Dissertations

The funnel plots previously presented suggest the presence of substantial bias in publication or analysis. If so, then non-significant results are less likely to be published; thus, unpublished dissertations may be less likely to have found statistical significance. Figure 5 highlights the unpublished dissertation experiments with funnel plots. 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. Table 4 provides these frequencies. The liberal counts assume independence of each entered effect size, while the conservative counts aggregate all effect sizes within each study. The aggregation in this latter counting strategy lead to three categories of studies: those that found significance on all outcomes, those that found significance on some outcomes, and those that found significance on no outcomes.

All tests were statistically significant. Across all paradigms, unpublished dissertations were much less likely to have found statistical significance than published studies (liberal and conservative tests, $p < .001$). Similarly, unpublished dissertations of all paradigms were far less likely to be included as best-practices than published studies (liberal test, $p < .001$; conservative test, $p = .003$). To the extent that these unpublished dissertations may reflect competent research less influenced by publication pressure, these results may be cause for concern. Similar results are also obtained when restricting these analyses to experiments: statistical significance, liberal test, $p < .001$, conservative test, $p = .001$; best-practices coding, liberal test, $p < .001$, conservative test, $p = .001$.

Meta-analytic effect size estimates were also drastically reduced within the set of experiments reported in unpublished dissertations. For aggressive affect, the random-effects estimate fell from $r = .22$ [.15, .29] in the full sample to $r = .02$ [-.10, .15] 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 $r = .20$ [.16, .24] in the full sample to $r = .13$ [.02, .24] in unpublished dissertations. These estimates should cause concern—they indicate that studies failing to find significant evidence for violent-game effects are more likely to go unpublished.

Discussion

Our findings indicate that there is strong evidence for substantial publication bias in the experimental studies of Anderson et al (2010). Publication bias was strongest among experiments with outcomes of aggressive affect and aggressive behavior, behavior being perhaps the most important outcome from pragmatic and theoretical considerations. Moreover, the bias was greatest in Anderson et al.'s "best-practices" subsets of these studies, which strikes us as

worrisome. Bias was relatively absent from experiments testing effects on aggressive cognition and physiological arousal. Publication bias is also absent from cross-sectional studies, which find correlations between violent game play and aggressive traits.

The core conclusion of this finding is that the experimental evidence for the violent-video game effects are less sound than has been presented. The original meta-analysis argued that all outcomes were statistically and practically significant. Yet, in our view, the degree of asymmetry in the funnel plots, especially those in Figures 1 and 2, makes it extremely difficult to trust this viewpoint. We find instead after adjustment that the effects of violent video games on aggressive behavior and affect in experiments are likely smaller than anticipated, and may be so small (r = .02–.15) as to be very challenging to detect in most experiments. Together, these analyses indicate that the evidence for causal effects of violent video games on aggressive outcomes, particularly aggressive affect and aggressive behavior, has been overstated. By way of comparison, we note that the evidence for experimental effects on aggressive cognitions (as opposed to aggressive affect or aggressive behavior) seems less affected by bias.

We urge some caution in interpreting the adjusted effect sizes. One problem is that the adjusted estimates are fairly divergent; for example, for aggressive affect in a best-practices experiment, they range from r = .24 to r = -.12. We see this as a consequence of the large degree of bias in this dataset. Were there little or no bias, the studies would have probative value, and the various meta-analytic adjustments would largely be in agreement. Instead, in certain subsets the bias is quite strong, and so the effect size estimate is perhaps too dependent on the modeling assumptions underlying the meta-analytic estimator. Suffice it to say that there are strong signs of bias, and we are confident only that the true effect sizes are smaller than naïve meta-analysis originally reported.

In contrast to the experimental literature, the cross-sectional literature seems relatively unbiased, and provides clear evidence of an association between violent video game use and aggressive thoughts, feelings, and behaviors. These correlations, however, cannot demonstrate causality, and may reflect a selection process (in that aggressive people may prefer violent games) or confounding by third variables (in that some other trait or process causes people to play violent video games and to behave aggressively). The longitudinal literature appears conflicted as to whether violent games cause aggressive behavior or aggressive behavior causes violent games (e.g., Breuer, Vogelgesang, Quandt, & Festl, 2015; Etchells, Gage, Rutherford, & Munafo, 2016; Willoughby, Adachi, & Good, 2012). Additionally, attempting to adjust for confounding variables such as sex appears to reduce effect sizes substantially (Anderson et al., 2010; Ferguson, 2015; Furuya-Kanamori & Doi, 2016; but see criticism from Boxer, Groves, & Docherty, 2015). Furthermore, we find considerable heterogeneity in effect sizes among cross-sectional studies; future research should determine why certain cross-sections find substantially larger or smaller effect sizes.

Theoretical Considerations

The current results suggest that theories of aggression may be weaker than previously thought. We consider theories of aggression and their proposed mechanisms, and what revisions may be necessary or in need of more careful testing.

The theoretical mechanisms used to explain violent video game effects, especially those in experimental situations, rely heavily upon the concept of "priming," in which exposure to some stimulus activates a related thought. One common theoretical perspective in social psychology is that merely thinking about a behavior increases the likelihood of performing that behavior (i.e., "ideomotor action", Bargh & Chartrand, 1999; James, 1890). In these theories, the

environment can cause one to think about certain behaviors. Thoughts are expected to be automatically and unavoidably influenced through perception, as perception is an automatic form of thought generally not subject to deliberate control. Thus, environmental cues are expected to activate thoughts, which in turn influence aggressive behavior; this "environment to perception to behavior" (Bargh & Chartrand, 1999, p. 468) chain is argued to operate smoothly, efficiently, automatically, and unconsciously. Furthermore, Bargh & Chartrand (1999) argue that these processes also influence affect, as the automatic evaluation of the environment activates affective states.

Theoretical accounts of the effects of violent video games propose a very similar process. Consider the General Aggression Model (GAM; Anderson & Bushman, 2002), perhaps the most frequently-applied theory of how and why violent video games would cause aggressive behavior. GAM theorizes that aggressive behavior is caused by internal states of arousal, aggressive affect, and aggressive cognitions. Violent video games are theorized to be an environmental influence, stimulating aggressive behavior by increasing arousal, priming the accessibility of aggressive thoughts, and inspiring aggressive feelings. Long-term effects are also sometimes explained in this framework; with repeated exposure across time, aggressive cognitions are expected to become "chronically accessible," causing the player's personality to become more aggressive (Bushman & Anderson, 2002).

The GAM and similar models in social psychology anticipate a clear and reliable experimental effect of violent media on aggressive behavior. The current results show the evidence for such an effect is overstated. Consequently, we think the evidence for the GAM is overstated as well. In this sense, our results fit in well with a groundswell of recent studies that question foundational findings in social priming and the "environment to perception to behavior"

chain. These studies include null effects of money primes on political attitudes (Rohrer, Pashler, & Harris, 2015), null effects of intelligence primes on general knowledge tasks (Shanks et al., 2013), and null effects of cleanliness primes on moral judgments (Johnson, Cheung, & Donnellan, 2014). These failures to replicate are also found in aggression research: DeWall and Bushman (2009) suggest that exposure to temperature-word primes increase hostile attributions, but a replication by McCarthy (2014) with a larger sample finds no such effect. Theories of aggression may need to reconsider whether incidental stimuli do in fact activate aggressive thoughts (a phenomenon that is itself quite ambiguously defined), and what manner of aggressive thoughts are necessary and sufficient to cause aggressive behavior.

Moderators of the effect. Recent commentaries have called for greater research emphasis on who is most likely to be influenced by violent-game effects, and thus, greater attention to potential moderators of violent-game effects (see, e.g., Krahé, in press). Our results suggest such efforts may be especially challenging. Publication bias in the previous research literature has likely obscured meaningful patterns of moderation from meta-analysis of multiple studies, and effects may be too small to reliably detect their moderators in single experimental studies. We elaborate on each point below.

First, we are concerned that publication bias may have the further side-effect of concealing patterns of moderation across studies in meta-analysis. Imagine that two series of studies measure the effects of violent games on aggressive behavior in two different populations. Imagine further that in one population, the effect size is moderate, and in the other population, the effect size is zero. Publication bias will favor results from the moderate-effect population and conceal results from the zero-effect population (or worse, encourage *p*-hacking such that the zero-effect population appears to have a moderate effect size). After publication bias,

theoretically-relevant moderators have been concealed by small-study effects. This may in part be why Anderson et al. (2010) had difficulty finding significant moderators in their analysis.

Second, we are concerned that effects may be too small to detect their moderators in single experiments. If the effects are indeed so small as we estimate, then individual studies will have to be quite large to detect the main effect, much less a significant moderator. For example, for aggressive behavior in a well-designed experiment, our largest estimate of the effect size recommended an adjustment from $r = .21$ to $r = .15$. Although this would seem to be a small adjustment, it is of substantial practical importance. Whereas the naïve estimate suggests a sample size of 136 is sufficient for 80% power in a one-tailed test, the PEESE estimate suggests that $n = 270$ is needed—a doubling of the sample size. The other adjustments all suggest that incredible sample sizes would be needed: *p*-curve, $r = .09$, $n = 759$; PET, $r = .07$, $n = 1{,}250$; *p*-uniform, $r = .02$, $n = 15{,}400$. To detect moderators of the effect would require still-larger sample sizes. In that sense, most experiments reporting moderators of violent-game effects are likely to be badly underpowered.

Nevertheless, many published studies report statistically-significant interactions of violent game content by individual differences such as trait anger or gender. We suspect that significant moderators are tested and discovered *post hoc* and may not be likely to replicate. 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. One of us has published such an interaction, trait anger × violent game exposure (Engelhardt, Bartholow, and Saults, 2011), and has experienced difficulty in replicating it

(Engelhardt, Mazurek, Hilgard, Rouder, and Bartholow, in prep). Another exploratory analysis of ours, claiming to find effects on cognitive control (Engelhardt, Hilgard, & Bartholow, 2015), was likely mistaken, as such "ego-depletion" effects could not be detected in a large-scale replication effort (Hagger et al., in press). The diversity of reported moderators and the infrequency of their replication suggest possible weaknesses in the literature of violent game effects.

The mediating role of aggressive affect? The General Aggression Model predicts that aggressive affect plays a mediating role between violent media and aggressive behavior. However, we did not find clear evidence that violent video games cause aggressive behavior in experiments; instead, studies indicated severe contamination by bias.

We suspect that this severe bias is caused by problems in inference that protect the hypothesis from falsification in this meta-analysis. Significant study results on aggressive affect are indeed interpreted as evidence that violent video games cause aggressive feelings. However, nonsignificant study results are not always interpreted as evidence that violent games do not cause aggressive feelings, and instead, are sometimes taken as evidence that the stimuli differ only in violent content and not in other confounding dimensions. The hypothesis can be changed after analyses to support the theory (Kerr, 1998).

If authors reported their null results as demonstrations of stimulus equivalence, they were excluded from meta-analysis. Anderson and colleagues (2010) state "Studies based on violent and nonviolent video games that have been preselected to [create equivalent affective states] 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" (page 156). Our concern is that stimuli may not have been truly *pre*selected

to create equivalent affective states. Preregistration of hypotheses and outcomes would prevent this ambiguity.

Furthermore, there seems to be little theoretical justification for why violent games should cause aggressive affect. Emphasis is usually placed instead on the activation of aggressive thoughts. Indeed, early reports specifically argued against violent-game effects on affect: Anderson and Dill (2000, p. 774) write "Violent content by itself, however, in the absence of another provocation, is likely to have little direct impact on affect." More recent reports hypothesize that arousal, aggressive cognition, and aggressive affect tend to co-activate; increases in aggressive cognitions, therefore, are hypothesized to cause increases in aggressive affect. Perhaps this mechanism could be tested directly.

Limitations

There are important limitations to the analyses we present. Although we are confident in the ability of funnel plots to detect small-study effects, we are less sure about the ability of our adjustments to provide accurate effect size estimates. We expect, at least, that they are reasonable estimates and may be closer to the truth than is the naïve estimate. Nonetheless, the statistical properties of these adjustments are not well understood, and the bias, efficiency, and robustness of these estimators 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 and *p*-uniform may under- or over-estimate effects in the context of *p*-hacking or heterogeneity.

These limitations of *p*-curve and *p*-uniform are particularly salient given concerns about the flexible analysis of the Competitive Reaction Time Task (Elson et al., 2014) and the presence of heterogeneity in certain analyses. It is possible that the underlying effect is substantial but our

estimates are biased in some direction by *p*-hacking in one or more studies, and it is possible that some *p*-curve/*p*-uniform estimates are too high due to heterogeneity.

Perhaps selection models (Vevea & Hedges, 1995) could provide a more effective and nuanced adjustment. We are particularly excited by the possibility of Bayesian selection methods (Guan & Vandekerckhove, 2016) that draw strength from reasonable prior information. 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 abstained from inspection of the partial effect sizes from the cross-sectional studies, as these can be challenging to synthesize properly. We have also abstained from inspection of longitudinal studies as there are not enough data points to permit a good estimate. It is possible, even likely, that there are small but detectable longitudinal effects of many hours of gameplay over time 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.

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 Greitemeyer and Mügge (2014) 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 a correction has withdrawn the 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.

Ways Forward

Meta-analysis, although exciting and informative, is fraught with difficult limitations. One productive way of avoiding these limitations is to conduct 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 (see, e.g., Hagger et al., in press; Matzke et al., 2015).

We suggest that those planning such a registered report consider the use of a modified-game paradigm (Elson, Bruer, Van Looy, Kneer, & Quandt, 2013; Elson & Quandt, 2014; Engelhardt, Hilgard, & Bartholow, 2015; Engelhardt, Mazurek, et al., 2015; Kneer et al., in

press). 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. 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. This would also allow meta-analyses of individual participant data, a particularly powerful and precise form of meta-analysis (see, e.g., Riley, Lambert, & Abo-Zaid, 2010), which would help to detect the moderators of the effect, if any.

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 in the published record (Lakens, Hilgard, & Staaks, 2016). Other researchers should be encouraged to inspect and reproduce meta-analyses. Meta-analyses that cannot be inspected or reproduced should be regarded with concern.

Summary

The research literature as analyzed by Anderson et al. (2010) seems to contain greater publication bias than their initial 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 favor statistically-significant results. Effects in experiments seem to be overestimated, particularly those of violent video game effects on aggressive behavior, which were estimated as being very close to zero. The theoretical insights into the causes and mechanisms of human aggression purportedly gained through this research program may enjoy less empirical support than originally reported. It may be appropriate to soften the rhetoric used in literature reviews and policy statements in order to respect this uncertainty and the possibility of smaller effect sizes than originally reported.

Rather than accept these adjusted 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. Our best hope for an accurate and informative hypothesis test rests upon an international, collaborative, and transparent research effort including proponents, skeptics, and disinterested third parties.

References

Aloe, A. M. (2014). An empirical investigation of partial effect sizes in meta-analysis of correlational data. *The Journal of General Psychology*, *141*, 47-64. DOI:10.1080/00221309.2013.853021

American Psychological Association Task Force on Violent Media. (2015). *Technical report on the review of the violent video game literature.* Retrieved from https://www.apa.org/news/press/releases/2015/08/technical-violent-games.pdf

Anderson, C. A., & Bushman, B. J. (2002). Human aggression. *Annual Review of Psychology, 53*, 27-51. DOI: 10.1146/annurev.psych.53.100901.135231

Anderson, C. A. & Dill, K. (2000). Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. *Journal of Personality and Social Psychology, 78*(4), 772-790. DOI: 10.1037/0022-3514.78.4.772

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. DOI:10.1037/a0018251

Ballard, M. E., & Wiest, J. R. (1996). Mortal Kombat: The effects of violent video game play on males' hostility and cardiovascular responding. Journal of Applied Social Psychology, 26, 717-730. DOI:10.1111/j.1559-1816.1996.tb02740.x

Bargh, J. A., & Chartrand, T. L. (1999) The unbearable automaticity of being. *American Psychologist, 54*(7), 462-479. DOI: 10.1037/0003-066X.54.7.462

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (2nd ed., pp. 221–235). New York, NY: Russell Sage Foundation.

Boxer, P., Groves, C. L. & Docherty, M. (2015). Video games do indeed influence children and adolescents' aggression, prosocial behavior, and academic performance: A clearer reading of Ferguson (2015). *Perspectives on Psychological Science, 10*(5), 671-673. DOI: 10.1177/1745691615592239

Breuer, J., Vogelgesang, J., Quandt, T., & Festl, R. (2015). Violent video games and physical aggression: Evidence for a selection effect among adolescents. *Psychology of Popular Media Culture, 4* (4), 305-328. DOI: 10.1037/ppm0000035

Bushman, B. J., & Anderson, C. A. (2002). Violent video games and hostile expectations: A test of the General Aggression Model. *Personality and Social Psychology Bulletin, 28*(12), 1679-1686. DOI: 10.1177/014616702237649

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. DOI:10.1037/a0018718

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*. DOI:10.3389/fpsyg.2014.00823

Council on Communications and Media. (2009). From the American Academy of Pediatrics: Policy statement – Media violence. *Pediatrics*, *124* (5), 1495-1503.

DeWall, C. N., & Bushman, B. J. (2009). Hot under the collar in a lukewarm environment: Words associated with hot temperature increase aggressive thoughts and hostile perceptions. *Journal of Experimental Social Psychology, 45* (4), 1045-1047. DOI: 10.1016/j.jesp.2009.05.003

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. DOI:10.1111/j.0006-341X.2000.00455.x

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ,* 315:629. DOI: http://dx.doi.org/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*. DOI:10.1037/ppm0000010

Elson, M., & Ferguson, C. J. (2014) Twenty-five years of research on violence in digital games and aggression: Empirical evidence, perspectives, and a debate gone astray. *European Psychologist*, *19*, 33-46. 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. DOI:10.1037/a0035569

Elson, M., & Quandt, T. (2014). Digital games in laboratory experiments: Controlling a complex stimulus through modding. *Psychology of Popular Media Culture*. DOI:10.1037/ppm0000033

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. DOI:10.1016/j.chb.2014.11.089

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*. DOI:10.1177/0956797615583038

Etchells, P. J., Gage, S. H., Rutherford, A. D., & Munafo, M. R. (2016). Prospective investigation of video game use in children and subsequent conduct disorder and depression using data from the Avon Longitudinal Study of Parents and Children. *PLoS One*. DOI:10.1371/journal.pone.0147732

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. DOI:10.1016/j.avb.2007.01.001

Ferguson, C. J. (2015). Do angry birds make for angry children? A meta-analysis of video game influences on children's and adolescents' aggression, mental health, prosocial behavior, and academic performance. *Perspectives on Psychological Science, 10*(5), 646-666. doi: 10.1177/1745691615592234

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. DOI:10.1016/j.jpeds.2008.11.033

Ferguson, C. J., & Kilburn, J. (2010). Much ado about nothing: The misestimation and overinterpretation of violent video game effects on eastern and western nations: Comment on Anderson et al. (2010). *Psychological Bulletin*, *136*, 174-178. DOI:10.1037/a0018566

Furuya-Kanamori, L., & Doi, S. A. R. (2016). Angry birds, angry children, and angry meta-analysts: A reanalysis. *Perspectives on Psychological Science, 11* (3), 408-414. doi: 10.1177/1745691616635599

Gervais, W. M. (2015, June 25). *Putting PET-PEESE to the test.* Blog post. Retrieved from http://willgervais.com/blog/2015/6/25/putting-pet-peese-to-the-test-1

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.

Greitemeyer, T., & Mügge, D. O. (2014). Video games do affect social outcomes: A meta-analytic review of the effects of violent and prosocial video game play. *Personality and Social Psychology Bulletin*, *40* (5), 578-589. DOI:10.1177/0146167213520459

Guan, M., & Vandekerckhove, J. (2016). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin and Review*, 23(1), 74-86. DOI:10.3758/s13423-015-0868-6

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Birt, A., Brand, R., … Cannon, T. (in press). A multi-lab pre-registered replication of the ego-depletion effect. *Perspectives on Psychological Science*.

Higgins, J. P. T. (2008). Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology, 37(5),* 1158-1160. DOI: 10.1093/ije/dyn204

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

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. DOI:10.1037/a0018567

Ioannidis, J. P. A., Trikalinos, T. A., & Zintzaras, E. (2006). Extreme between-study homogeneity in meta-analyses could offer useful insights. *Journal of Clinical Epidemiology, 59*(10), 1023-1032. DOI: 10.1016/j.jclinepi.2006.02.013

James, W. (1890) *The principles of psychology (Vol. 2).* New York: Holt.

Johnson, D. J., Cheung, F., & Donnellan, M. B. (2014). Does cleanliness influence moral judgments? A direct replication of Schnall, Benton, and Harvey (2008). *Social Psychology, 45,* 209-215. DOI: 10.1027/1864-9335/a000186

Kerr, N. L. (1998) HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review, 2* (3), 196-217. DOI: 10.1207/s15327957pspr0203_4

Kneer, J., Elson, M., & Knapp, F. (2016). 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, 54,* 142-148. DOI:10.1016/j.chb.2015.07.034

Krahé, B. (in press). Violent media effects on aggression: A commentary from a cross-cultural perspective. DOI: 10.1111/asap.12107

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BioMed Central*. Retrieved from http://tinyurl.com/LakensHilgardStaaks

Lau, J., Ioannidis, J. P. A., Terrin, N., Schmid, C. H., & Olkin, I. (2006). The case of the misleading funnel plot. *BMJ, 333*. DOI:0.1136/bmj.333.7568.597

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E.-J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Psychology: General, 144* (1), e1-e15. DOI:10.1037/xge0000038

McCarthy, R. J. (2014). Close replication attempts of the heat priming-hostile perception effect. *Journal of Experimental Social Psychology, 54,* 165-169. DOI: 10.1016/j.jesp.2014.04.014

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology, 9*. DOI:10.1186/1471-2288-9-2

O'Boyle, E. H., Jr., Banks, G. C., & Gonzalez-Mule, E. (2014). The chrysalis effect: How ugly intitial results metamorphosize into beautiful articles. *Journal of Management*. DOI:10.1177/0149206314527133

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. 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 DOI:10.1037/a0034820

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *BMJ, 340,* DOI: 10.1136/bmj.c221

Rohrer, D., Pashler, H., & Harris, C. R. (2015). Do subtle reminders of money change people's political views? *Journal of Experimental Psychology: General*, *144* (4), e73-e85. DOI: 10.1037/xge0000058

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., Kavvadia, F., & Moore, C. (2013). Priming intelligent behavior: An elusive phenomenon. *PLoS ONE*. DOI: 10.1371/journal.pone.0056515

Sigurdsson, J. F., Gudjonsson, G. H., Bragason, A. V., Kristjansdottir, E., & Sigfusdottir, I. D. (2006). The role of violent cognition in the relationship between personality and the involvement in violent films and computer games. *Personality and Individual Differences*, *41*, 381-392. DOI:10.1016/j.paid.2006.02.006

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. DOI: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. DOI:10.1177/1745691614553988

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

Sterne, J. A. C., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, *53* (11), 1119-1129. DOI:10.1016/S0895-4356(00)00242-0

Tear, M. J., & Nielsen, M. (2014). Video games and prosocial behavior: A study of the effects of non-violent, violent and ultra-violent gameplay. *Computers in Human Behavior*, *41*, 8-13. DOI:10.1016/j.chb.2014.09.002

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, *22*, 2113-2126. DOI:10.1002/sim.1461

Urashima, M., & Suzuki, K. (2003). Konpyuuta gemu ga kodomo no koudou ni oyobosu eikyo [the effects of playing with computer games on children's behavior]. *Journal of Child Health*, *50*, 50-56.

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (in press). Conducting meta-analyses based on *p*-values: Reservations and recommendations for applying *p*-uniform and *p*-curve. *Perspectives on Psychological Science*. Retrieved from http://www.metaresearch.nl/wp-content/uploads/2016/04/Preprint_VanAert_Wicherts_VanAssen16.pdf

van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, *20*, 293-309. DOI:10.1037/met0000025

Viechtbauer, W. (2010). Conducting meta-analyses in R with the `metafor` package. *Journal of Statistical Software, 36* (3). Retrieved from http://www.jstatsoft.org/v36/i03/

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, 60*, 419-435. DOI:10.1007/BF02294384

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. DOI:10.1037/a0026046

Table 1
Naïve effect-size estimates

Setting	Group	*k*	*N*	Fixed	Random	I^2
			Aggressive Affect			
Experiment	Best	18	1318	.29 [.24, .34]	.34 [.24, .42]	66 [45, 91]
Experiment	Full	34	2879	.17 [.14, .21]	.22 [.15, .29]	72 [61, 89]
Cross-Section	Best	7	4348	.10 [.07, .13]	.10 [.05, .16]	65 [12, 96]
Cross-Section	Full	14	9811	.15 [.13, .17]	.16 [.08, .24]	93 [87, 98]
			Aggressive Behavior			
Experiment	Best	23	2413	.21 [.17, .25]	.21 [.17, .25]	4 [0, 17]
Experiment	Full	39	3328	.17 [.14, .20]	.17 [.14, .20]	0 [0, 7]
Cross-Section	Best	22	12391	.29 [.27, .30]	.30 [.25, .35]	88 [77, 93]
Cross-Section	Full	37	29113	.21 [.20, .22]	.24 [.21, .28]	91 [84, 94]
			Aggressive Cognition			
Experiment	Best	24	2887	.22 [.18, .25]	.22 [.18, .27]	35 [0, 70]
Experiment	Full	40	4073.5	.20 [.17, .23]	.20 [.16, .24]	27 [0, 67]
Cross-Section	Best	17	7997	.21 [.19, .23]	.21 [.15, .28]	87 [75, 94]
Cross-Section	Full	22	13012	.18 [.17, .20]	.21 [.15, .27]	91 [83, 95]
			Physiological Arousal			
Experiment	Best	11	833	.20 [.13, .26]	.21 [.11, .31]	50 [0, 80]
Experiment	Full	24	1770	.14 [.09, .18]	.15 [.09, .21]	35 [0, 71]

Note: K = number of studies; *N* = total N across studies. All effect sizes in Pearson *r* *w*ith 95% confidence intervals.

Table 2

Tests for bias and small-study effects.

Setting	Group	b_{Egger}	SE(b_{Egger})	p_{Egger}	$p_{p\text{-}uniform}$
		Aggressive Affect			
Experiment	Best	3.667	0.78	**< .001**	0.201
Experiment	Full	2.635	0.737	**< .001**	0.861
Cross-Section	Best	-	-	-	-
Cross-Section	Full	0.123	1.883	0.948	0.661
		Aggressive Behavior			
Experiment	Best	1.537	0.549	**0.005**	**0.002**
Experiment	Full	0.451	0.39	0.248	**0.009**
Cross-Section	Best	1.163	0.789	0.140	0.752
Cross-Section	Full	1.326	0.589	**0.024**	0.900
		Aggressive Cognition			
Experiment	Best	1.372	0.761	0.071	0.684
Experiment	Full	0.883	0.544	0.104	0.814
Cross-Section	Best	-0.447	1.469	0.761	0.628
Cross-Section	Full	0.334	1.366	0.807	0.544
		Physiological Arousal			
Experiment	Best	0.137	1.22	0.911	0.797
Experiment	Full	1.295	0.714	0.070	0.930

Note: One analysis omitted for insufficient number of studies. Bold text highlights tests significant at the .05 level.

Table 3
Adjusted effect-size estimates.

		PET	I^2_{PET}	PEESE	I^2_{PEESE}	*p*-uniform	*p*-curve
				Aggressive Affect			
Experiment	Best	-.12 [-.29, .06]	0 [0, 83]	.14 [.06, .23]	0 [0, 86]	.24 [.08, .36]	.21
Experiment	Full	-.10 [-.27, .08]	58 [44, 85]	.08 [-.02, .18]	60 [47, 86]	.24 [.11, .35]	.20
Cross-Section	Best	-	-	-	-	-	-
Cross-Section	Full	.16 [-.04, .35]	94 [88, 98]	.17 [.04, .29]	94 [88, 98]	.16 [.12, .24]	.16
				Aggressive Behavior			
Experiment	Best	.07 [-.04, .18]	0 [*]	.15 [.09, .21]	0 [*]	.02 [-.23, .15]	.09
Experiment	Full	**.13 [.04, .21]**	0 [0, 61]	.15 [.10, .20]	0 [0, 61]	.02 [-.23, .15]	.08
Cross-Section	Best	**.29 [.16, .41]**	87 [77, 94]	.30 [.23, .37]	88 [78, 94]	.27 [.24, .31]	.28
Cross-Section	Full	**.21 [.12, .28]**	90 [84, 94]	.24 [.19, .28]	91 [85, 94]	.22 [.19, .25]	.23
				Aggressive Cognition			
Experiment	Best	.10 [-.05, .24]	33 [0, 65]	.18 [.11, .24]	32 [0, 65]	.24 [.15, .31]	.19
Experiment	Full	**.11 [.00, .22]**	29 [0, 64]	.16 [.10, .21]	27 [0, 62]	.24 [.14, .32]	.19
Cross-Section	Best	**.24 [.07, .39]**	88 [76, 95]	.23 [.14, .32]	88 [77, 95]	.18 [.14, .22]	.19
Cross-Section	Full	**.20 [.05, .33]**	91 [84, 96]	.22 [.13, .30]	91 [84, 96]	.16 [.13, .20]	.18
				Physiological Arousal			
Experiment	Best	.19 [-.12, .47]	53 [0, 83]	.21 [.04, .37]	54 [0, 84]	.26 [.08, .37]	.28
Experiment	Full	-.01 [-.18, .17]	31 [0, 66]	.09 [.00, .17]	32 [0, 65]	.26 [.08, .37]	.28

Note: K = number of studies; *N* = total N across studies. * Confidence interval on I^2 consists of the null/empty set due to highly homogeneous data. One analysis omitted for insufficient number of studies. Bold text indicates where the 95% CI of the PET estimate excludes zero, suggesting that the underlying effect is nonzero and that PEESE should be favored over PET. All effect sizes in Pearson *r*.

Table 4

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

Liberal coding scheme		
	Statistical significance	
Publication format	Yes	No
Unpublished Dissertation	4	30
Other	201	125
	Labeled Best Practices	
Publication format	Yes	No
Unpublished Dissertation	4	30
Other	208	118

Conservative coding scheme			
	Statistical significance		
Publication format	All outcomes	Some outcomes	No outcomes
Unpublished Dissertation	2	2	14
Other	73	34	29
	Labeled Best Practices		
Publication format	Yes	No	
Unpublished Dissertation	3	15	
Other	83	62	

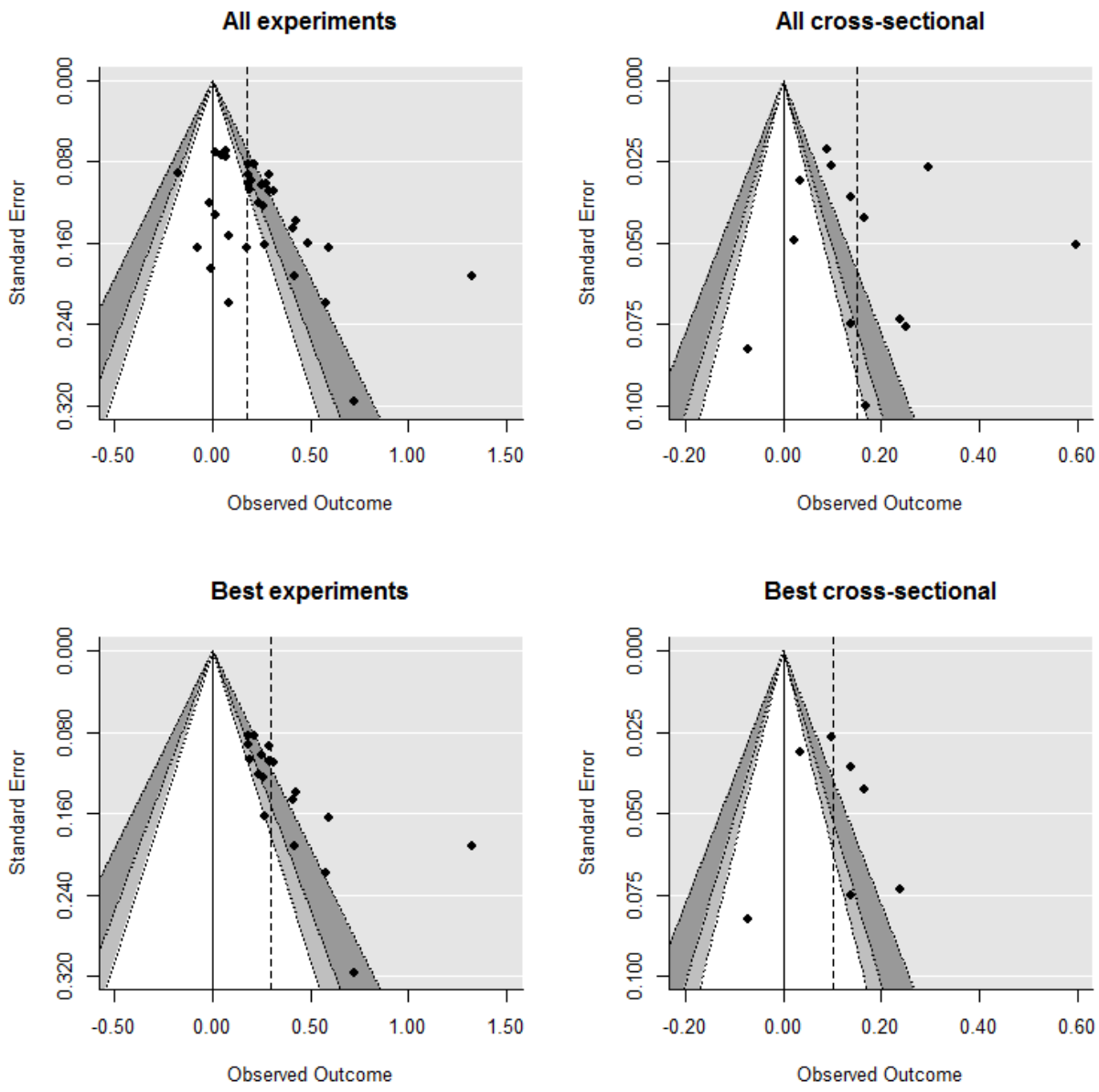

Figure 1. Funnel plot of studies of aggressive affect with shaded contours for .05 < *p* < .10 (light grey) and .01 < *p* < .05 (dark grey). Application of best-practices criteria seems to emphasize statistical significance, and a knot of experiments just reach statistical significance. One best-practices experiment (Ballard & Wiest, 1996) finds an implausibly large effect (*z* = 1.33), and one not-best-practices cross-sectional study appears to be an outlier (Urashima & Suzuki, 2003, *z* = 0.60)

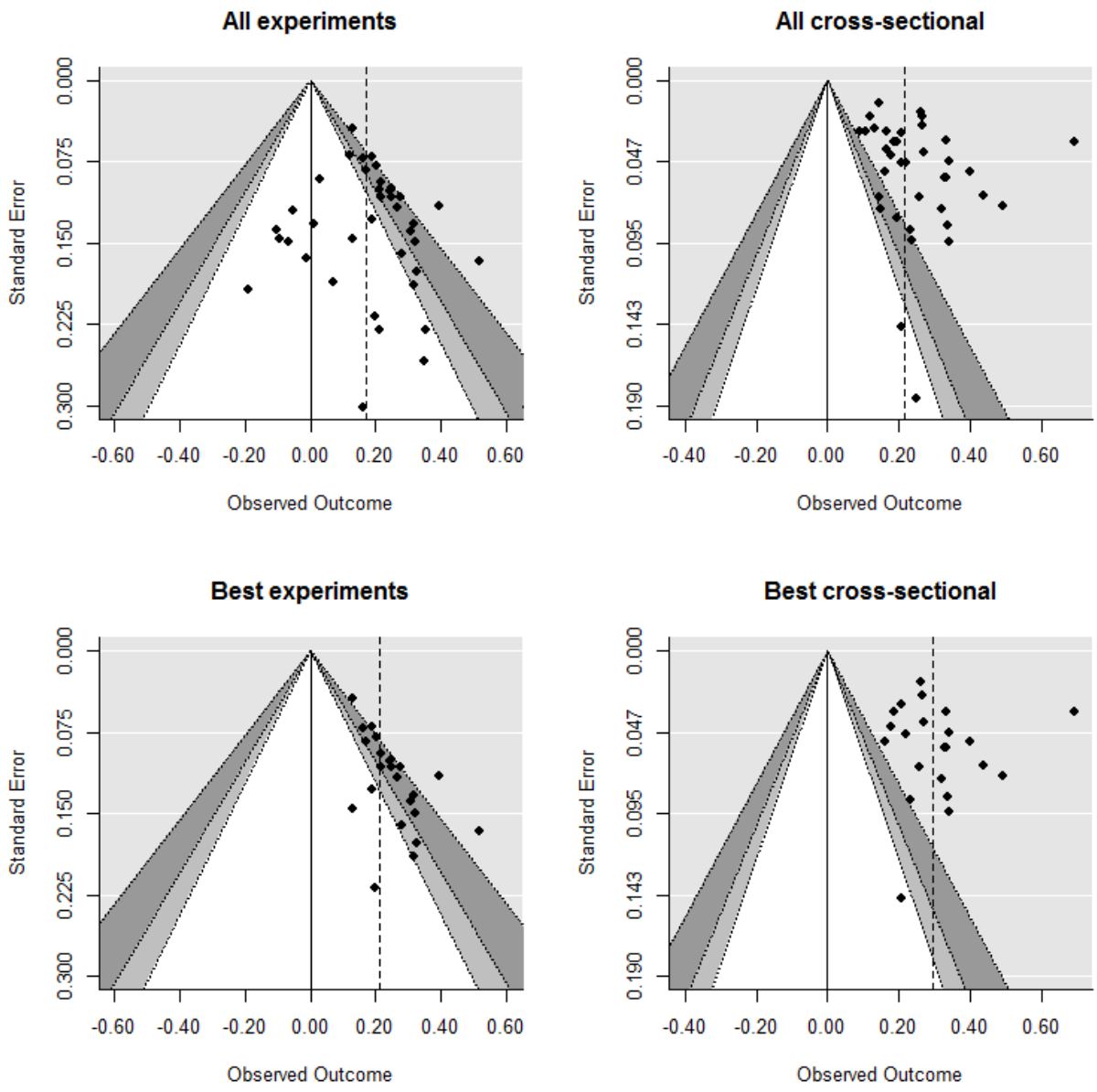

Figure 2. Funnel plot of studies of aggressive behavior with shaded contours for .05 < *p* < .10 (light grey) and .01 < *p* < .05 (dark grey). Application of best-practices criteria seems to emphasize statistical significance, and a knot of experiments just reach statistical significance. Again, application of best-practices criteria favors experiments finding statistical significance. One best-practices cross-sectional study appears to be an outlier (Matsuzaki, Watanabe, & Satou, 2004, *z* = .69).

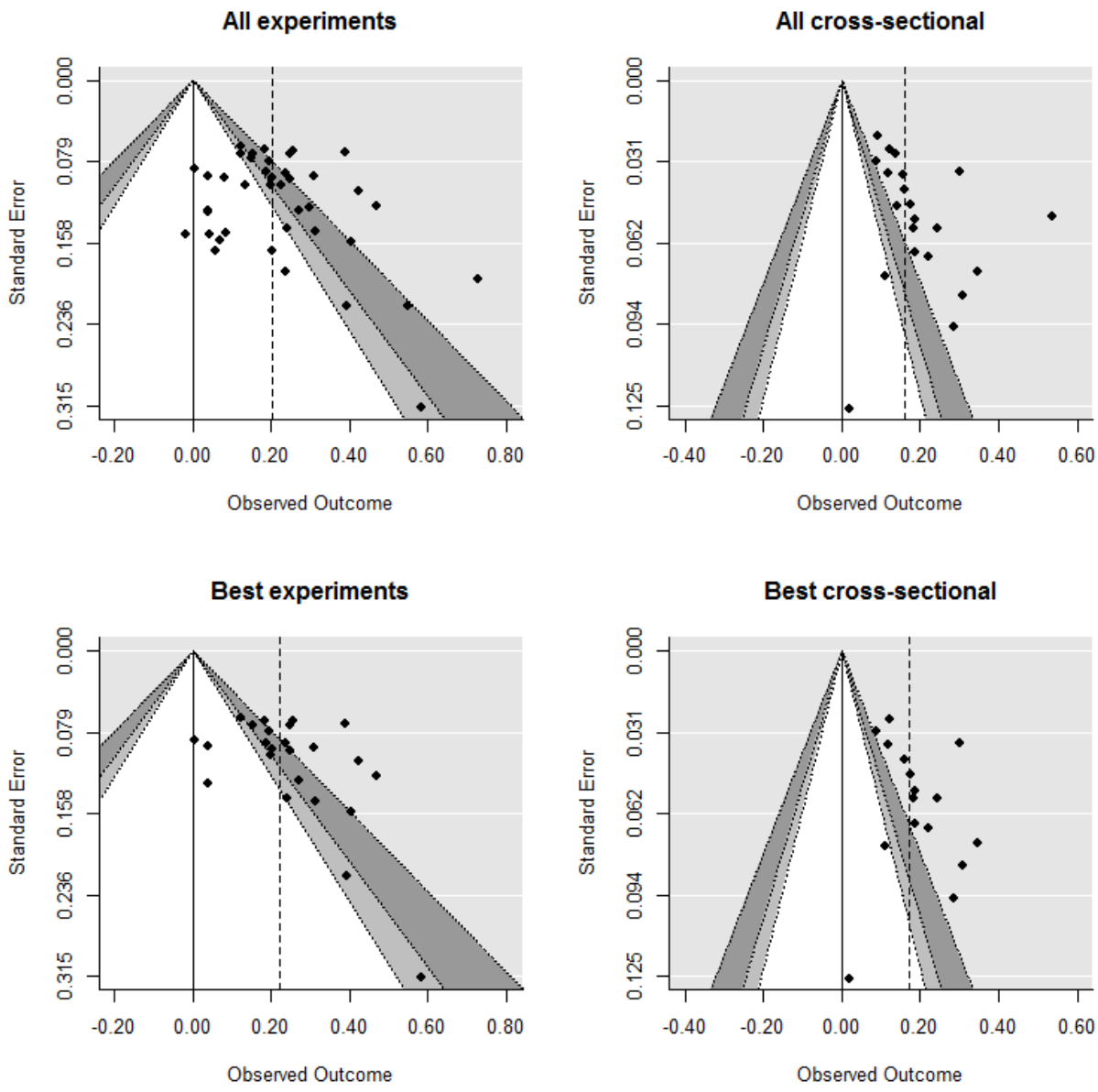

Figure 3. Funnel plot of studies of aggressive cognition with shaded contours for .05 < *p* < .10 (light grey) and .01 < *p* < .05 (dark grey). Results appear moderately heterogeneous, but not particularly contaminated by bias. One not-best-practices cross-sectional study may be an outlier (Sigurdsson, Gudjonsson, Bragason, Kristjansdottir, & Sigfusdottir, 2006, *z* = 0.53).

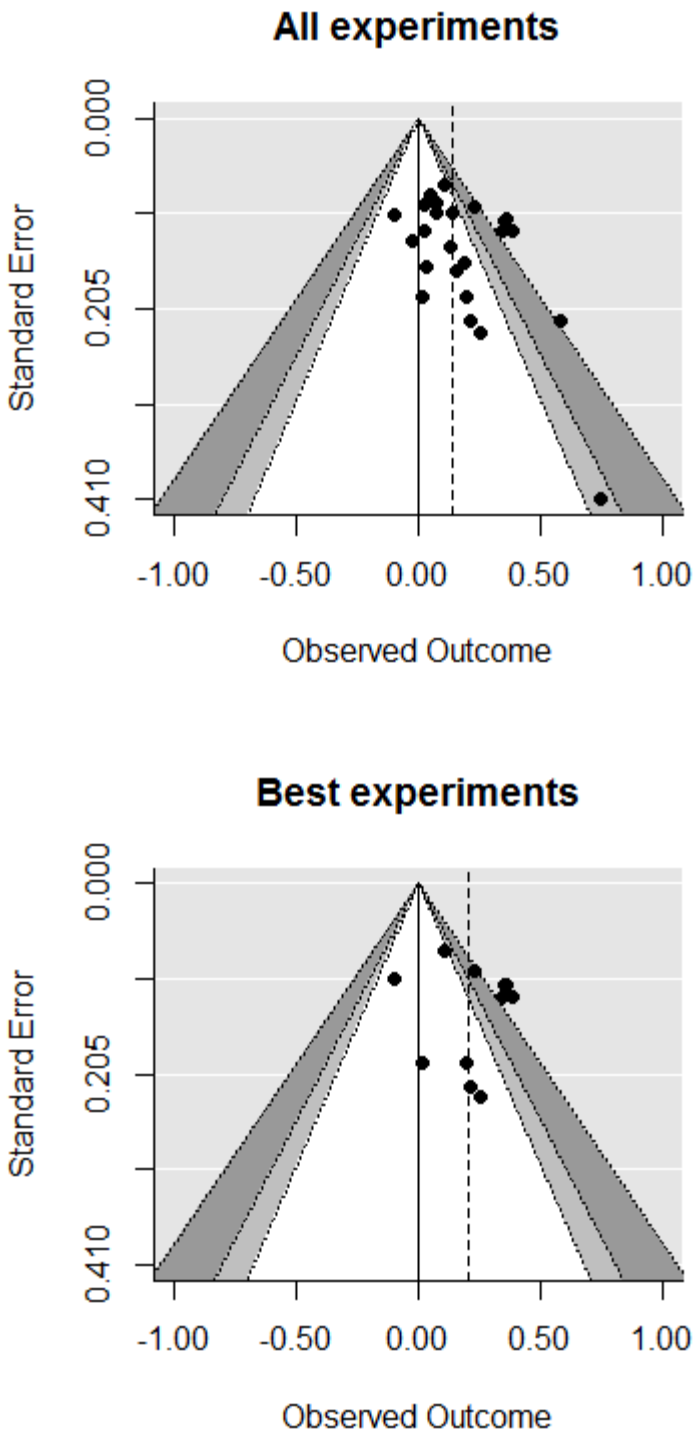

Figure 4. Funnel plot of studies of physiological arousal with shaded contours for .05 < *p* < .10 (light grey) and .01 < *p* < .05 (dark grey). Results do not appear to be systematically contaminated by bias.

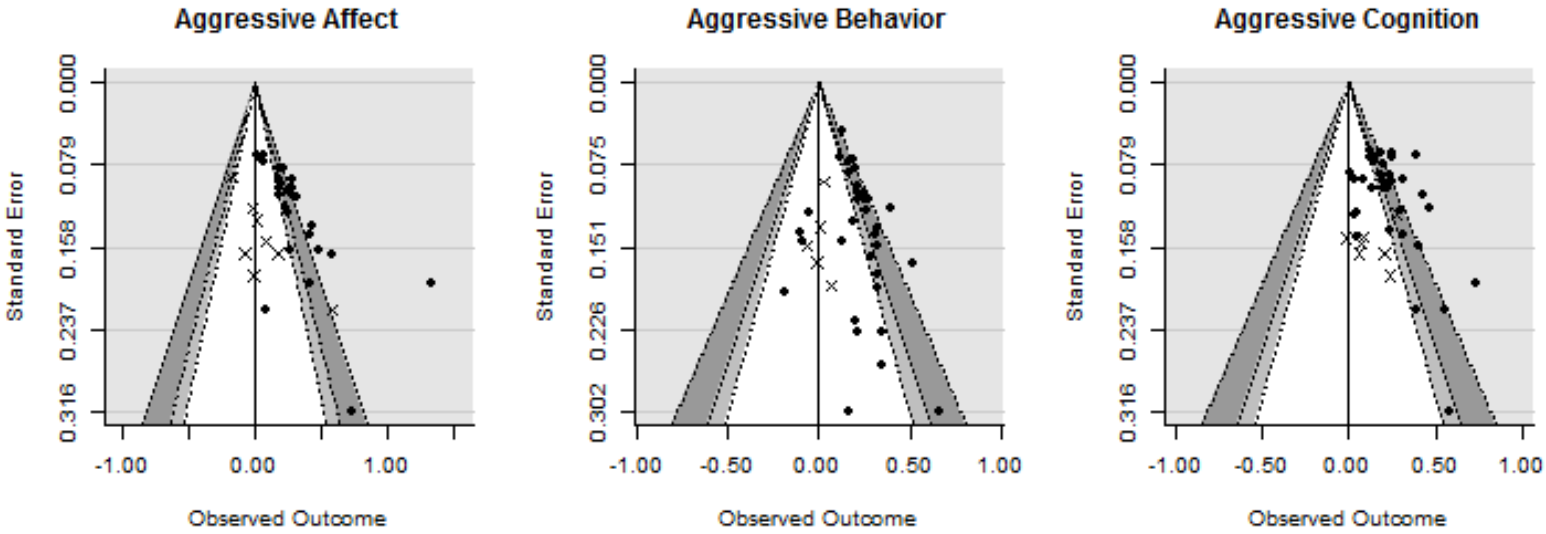

Figure 5. Funnel plots of all experiments of aggressive affect, behavior, and cognition. Dissertations not presented in any further publication format are indicated with Xs, while all other publication styles (e.g., journal articles, book chapters, conference proceedings) are indicated with filled dots. Shaded contours represent two-tailed *p*-values between .10 and .05 (light grey) and between .05 and .01 (dark grey). 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.