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

Joseph Hilgard, Christopher R. Engelhardt, Bruce D. Bartholow, and Jeffrey N. Rouder

University of Missouri

Author Note

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

Abstract

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

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

Do violent games make their players more aggressive? Despite decades of research and hundreds of studies, there remains scientific debate. Results have been aggregated in several meta-analyses, the most-cited and most comprehensive of which (C. A. Anderson et al., 2010) claimed decisive evidence for effects on aggressive thoughts, feelings, and behaviors in experimental, cross-sectional, and longitudinal research designs. This impressive volume of converging research findings has been hailed by some as decisive evidence (**???**). Policy statements from a number of professional organizations (e.g., on Communications & Media, 2009) reflect this perspective, urging the public of the considerable evidence demonstrating harmful outcomes of violent media use.

Despite this meta-analysis, there are still skeptics of causal effects of violent video games on aggressive outcomes. One point of skepticism is that the actual effect size is closer to zero. Supporting this point, meta-analyses conducted by skeptics estimate smaller effects (Sherry, 2001; **?**; **?**; **?**), and some experiments by skeptics fail to detect significant effects of violent game content (Adachi & Willoughby, 2011; Elson, Bruer, Van Looy, Kneer, & Quandt, 2013; Ferguson et al., 2008; Valadez & Ferguson, 2012). However, these meta-analyses are sometimes criticized for containing fewer studies than the C. A. Anderson et al. (2010) meta-analysis (**?**, see, e.g., a response by), and the weight of evidence in nonsignificant experimental results tends to vary (**?**).

Another point of skepticism is that the obtained effects may be statistically robust, but the interpretation is flawed. One argument is that the effects observed in experiments are due to confounds, not violent content (**?**). In experimental research, these confounds would include differences between video games in dimensions other than violent content. In correlational and longitudinal research, these confounds are proposed also to include dispositional features that may attract players to both violent games and aggressive behavior over time. Others argue that the public impact of the research has been overstated. This argument questions the external validity of laboratory measures and

posits that causal changes observed in laboratory measures may not reflect real-world causal changes in real-world aggressive behavior (**?**).

However, the debate is not limited to theoretical and methodological concerns. Skeptics have suggested that the literature is contaminated by biases in analysis and report that make the evidence appear stronger than it is. For example, **?** suggests that studies that do not find significant effects are less likely to be published. If this is the case, then meta-analysis of the published data would systematically overestimate the effect, observing studies that estimate larger effects but not observing studies that estimate smaller or even negative effects. Similarly, others have suggested that obtained study data have been flexibly analyzed until the desired research conclusion was reached (**?**). Such flexible analysis would bias the results of individual studies, nudging their effect sizes until they were large enough to reach statistical significance. In aggregate, then, the sum of these biased studies would itself be biased, again overestimating the true effect size. If either of these are the case, then the extant data may not permit an appropriate hypothesis test, however overwhelming the evidence may otherwise seem to be. In the presence of bias in publication or analysis, the effect of violent games will be overestimated.

Another point of contention has been the application of "best-practices criteria" to studies gathered for meta-analysis. In their meta-analysis, C. A. Anderson et al. (2010) collected all available studies, then applied a set of "best-practices criteria" to separate these into what they argued were and were not appropriate tests of the research hypothesis. The authors reported that studies that had been performed according to these criteria found larger effects of violent games than did studies that had not. It has been argued, however, that these criteria were vague in definition and inconsistent in application (**??**). It is implied that the inconsistency in application was motivated, with the meta-analysts selecting hypothesis-confirming studies as being best-practices studies while discarding studies with non-significant results as being not-best-practices studies.

In the present manuscript, we inspect the strength of the available literature by

revisiting the meta-analysis presented by C. A. Anderson et al. (2010). In that manuscript, authors applied a trim-and-fill procedure (**?**) to inspect and adjust for the presence of bias. However, the trim-and-fill procedure is understood to be flawed, having assumptions that are likely to be unmet in actual practice. It is expected to under-correct in the presence of bias and over-correct in the absence of bias (**??**). New meta-analytic techniques have since been developed that promise greater accuracy than trim-and-fill. We apply two of these, *p*-curve and PET-PEESE, to the dataset provided by Anderson and colleagues, reporting adjusted effect sizes and new inferences.

Is the Debate Concluded?

> Despite thousands of research studies on media effects, many people simply refuse to believe them. Some academics may contribute to this because they like to "buck the establishment," which is an easy way to promote themselves and their research. Of course, many people still believe that President Obama wasn't born in the United States, President Kennedy wasn't assassinated, men didn't walk on the moon, and the Holocaust didn't occur. (**?**, p. 572)

In the past few years, there has been a change in the tone of the debate surrounding violent video game effects. Some proponents are sufficiently convinced of the effects that any remaining skepticism seems to be due to unfathomable stubbornness. Because these proponents consider the research question settled, they have moved to novel research topics. Some study conceptual extensions of the basic phenomenon to include prosocial effects of prosocial video games. Others study science denial and biased assimilation among those motivated to distrust violent games research.

This latter topic has begun to expand into a line of science-communication research exploring ways to understand skepticism about violent-game effects. For example, Nauroth, Gollwitzer, Bender, and Rothmund (2014) present evidence that gamers are themselves particularly resistant to evidence of negative effects of video games. Greitemeyer (2014)

similarly finds that readers favor studies of violent-game effects that confirm their beliefs.

Strategies are now being developed to make violent media research more convincing and actionable. One such strategy is the proposal that researchers "advance the debate" by speaking directly to the public to avoid skeptics within academia (**?**). Claims of consensus have been advanced (**?**), and attempts have been made to separate "true media violence scientists" who believe in the effect from less-expert sources whose conflicts of interest prevent honest evaluation of the available evidence (**?**). It has been suggested that the next challenge for violent-games research is not to better understand the phenomenon, but rather, to foster belief in the phenomenon among both the wider scientific community and the laity, perhaps in part by excluding skeptics from public debate (**?**).

Whether this is an appropriate course of action depends on the strength of evidence. If the evidence is incontrovertible, then skeptics may be misleading their audiences by refusing to update their beliefs in light of research findings. One could debate the nature of the effect and the validity of manipulations and measures, but one could not argue that the effect does not exist. On the other hand, if the evidence is flawed, then proponents risk stifling research and debate where it is most needed. A thorough and conservative inspection of the evidence, then, is of vital importance.

Publication Bias and Small-Study Effects

In recent years, psychology has experienced a crisis of confidence as researchers realize that many published research findings may be false. Using statistical techniques and reporting standards typical of social psychology, researchers have been able to provide experimental evidence for impossible phenomena such as extra-sensory precogition (psi Bem, 2011) and a song that makes its listeners younger (Simmons, Nelson, & Simonsohn, 2011). Critics have pointed out that hypothesis-confirming results appear in the literature much more frequently than would be expected given reasonable estimates of statistical power. It has even been suggested that the current "publish or perish" reward structure of

academia encourages capitalization on Type I error, encouraging researchers to publish many studies with poor predictive value rather than publish few studies with substantial predictive power (**?**). In this light, one might expect that there could be bias in violent games research, as there is in so many other disciplines.

Two processes may contribute to such research bias. The first, publication bias, is the phenomenon that studies with statistically significant (i.e., $p < .05$) findings are more likely to be submitted and accepted for publication. Publication bias is a problem that contributes to the overestimation of effect sizes and the propagation of Type I error. It is an especially pernicious problem for meta-analysis, as the selective reporting of studies that "work" (i.e., attain significance) leads to an overestimated effect size and may lead to conclusions of statistically and practically significant effects when there are none. The error introduced by publication bias is larger when research studies are underpowered, as only the studies that overestimate the effect dramatically are able to reach the threshold of statistical significance.

The other process is called by many names: flexible analysis, questionable research practices, p-hacking. These names refer to biased research practices that increase the likelihood of finding significant effects by increasing Type I error rates. One such practice is the inspection of many statistical tests and the presentation of only the significant ones. For example, one might collect several study outcomes but report only the one that showed significant differences, censoring the non-significant outcomes from report. Similarly, one might collect several treatment conditions but censor from report those conditions whose outcomes do not support the hypothesis. One could go "moderator munging", exploring several moderators until a significant interaction is found. Covariates could be added or removed from the model until the desired relationship becomes significant. Observations might be labeled as outliers and excluded not for their leverage, but for whether they support or oppose the hypothesized relationship. One particularly subtle form of flexible analysis is "sampling to a foregone conclusion," in which the p-value is repeatedly

inspected and additional data is collected until the p-value reaches the necessary threshold. While such sequential analyses can be appropriate and efficient in preregistered research plans, they have historically been used in an *ad hoc* fashion that inflates Type I error rates and effect size estimates.

Because these two problems are typical in research, many meta-analytic techniques have been developed to detect and adjust for research bias. The application of such techniques are a vital part of meta-analytic practice. Additionally, because new techniques are continuously being developed, each promising potential improvements in accuracy, it may be helpful to revisit previous meta-analyses and apply new techniques for detecting and adjusting for publication bias (**?**).

In the C. A. Anderson et al. (2010) meta-analysis, the authors applied one popular technique, the trim-and-fill procedure, to suggest bias-adjusted effect size estimates. This procedure yielded minimally-adjusted estimates, suggesting minimal bias. However, there are other ways to adjust for bias in meta-analysis. In the following section, we review some meta-analytic techniques for detecting and adjusting for bias, describing their properties, strengths, and weaknesses.

Egger's regression test. One simple test for research bias is Egger's regression test (Egger, 1997). This test inspects the relationship between effect size and precision (or sample size) in reported studies. Because sample size does not typically cause effect size, an unbiased research literature is expected to have no relationship between effect size and precision. However, if studies must attain statistical significance to be published, such a relationship will be observed. Small-sample studies require large observed effect sizes to reach statistical significance, while large-sample studies can reach statistical significance with smaller observed effect sizes. Thus, in the presence of publication bias, there is an inverse relationship between effect size and precision. Egger's regression test inspects the degree and statistical significance of this relationship.

Note that, in some cases, sample size and effect size may be correlated for reasons

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

One weakness of Egger's regression test is that, while it can detect bias, it does not suggest a bias-adjusted effect size. Thus, it is not possible to assess whether the meta-analytic estimate reflects a likely null value or some non-null but inflated value. The test has also been demonstrated to have poor statistical power, limiting the strength of conclusions that can be drawn through application of the test.

Egger's regression test has been used repeatedly by skeptics to look for publication bias (e.g., Ferguson & Kilburn, 2009; **?**), but was not reported in the C. A. Anderson et al. (2010) meta-analysis. Thus, while Anderson and colleagues argue that their analysis contains minimal publication bias, an Egger's regression test might have disagreed.

Funnel plots. Because research bias is one potential cause of small-study effects, it is often useful to visually inspect meta-analytic data for small-study effects. The relationship between observed effect size and precision is often represented for this purpose in a funnel plot. In a funnel plot, effect size is plotted on the x-axis and precision on the y-axis. In the absence of small-study effects or heterogeneity, study results will form a symmetrical funnel shape, displaying substantial variance when sampling error is large but narrowing to a precise estimate when sampling error is small. Thus, when research is not contaminated by bias, some small-sample studies are expected to find null or even negative results due to sampling error. The funnel should fill evenly.

However, when there are small-study effects, the funnel plot is no longer symmetrical.

In the case of publication bias, studies are missing from the lower portion of the funnel where results would not be statistically significant. Funnel-plot asymmetry can also be caused by flexible analysis and reporting. When samples are collected until a desired *p*-value is attained, studies will move up and to the right of the funnel. When subgroups or experimental subgroups are dropped from report to highlight only a subgroup in which statistical significance was found, studies will move down and to the right. When outcomes are censored from report to highlight only the significant outcomes, studies will move to the right of the funnel.

Again, funnel plots have been presented by skeptics (e.g., **?**), but the C. A. Anderson et al. (2010) meta-analysis did not provide any funnel plots. This makes it difficult for readers to appraise the strength of the data, inspect the distribution of study results, and determine whether the naive and trim-and-fill effect size estimates might be influenced by outliers.

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

However intuitive, this is not an effective adjustment for bias, as the assumptions of trim-and-fill are unlikely to be met. Studies are not likely to be censored on the basis of the effect size, but rather, on the basis of their statistical significance. Accordingly, it is argued that trim-and-fill does a poor job of providing an adjusted effect size, adjusting too much when there is no bias and adjusting too little when there is bias (**??**). Others are skeptical of trim-and-fill's imputation of studies.

Thus, trim-and-fill is most commonly suggested as a form of sensitivity analysis rather than a serious estimate of the unbiased effect size. When the naive meta-analytic estimate and the trim-and-fill-adjusted estimate differ only slightly, it is suggested that the research is largely unbiased. C. A. Anderson et al. (2010) applied trim and fill in their meta-analysis as the only attempt to detect and adjust for small-study effects. Trim-and-fill yielded only slightly-adjusted effect sizes, and so the authors concluded minimal research bias. Some have characterized this as an extensive test for publication bias (**?**, pg. 51) despite the weaknesses of the trim-and-fill procedure and the absence of funnel plots.

PET-PEESE meta-regression. A promising new tool in the detection of and adjustment for bias is meta-regression. Like Egger's test, meta-regression techniques for publication bias consider the relationship between effect size and precision. Under publication bias, larger samples yield smaller effects. Again, because sample size does not typically cause effect size, such a relationship between sample size and effect size suggests that studies were censored when not attaining statistical significance or that studies were flexibly analyzed in order to attain statistical significance.

PET-PEESE meta-regression (**?**) uses the relationship between precision and effect size to estimate the underlying effect. It does this in two steps: Precision-Effect Test (PET) and Precision-Effect Estimate with Standard Error (PEESE).

In PET, a weighted *linear* regression is fit to describe the relationship between effect size and precision, then extrapolates to estimate what the "true effect" would be in a hypothetical study with perfect precision. This true effect corresponds to the estimated intercept in the metaregression equation describing effect size as a function of precision. That is, the intercept represents the estimated effect size after partialing out the linear effect of sample size on effect size.

When there is no true effect, published studies tend to lie on the boundary between statistical significance and nonsignificance, forming a linear relationship between sample size and precision. Thus, PET performs well at estimating effects when the null hypothesis

is roughly true. However, when there is a true effect, small studies will be censored by publication bias, but most large studies will find statistical significance and be unaffected by bias. PET will fail to model this nuance and risks underestimating the size of true effects.

A second meta-regression estimator, PEESE, is intended to address this problem. PEESE fits a weighted *quadratic* relationship between effect size and precision. The resulting curve models bias as being stronger in the lower part of the funnel but reduced as the studies become better-powered and less subject to bias. PEESE is thought to perform well in estimating nonzero effects, but risks overestimating the size of null effects.

The PET-PEESE predictor is intended to address the complementary strengths and weaknesses of the two estimators by combining them in a single conditional estimator. First, PET is applied and the significance of its adjusted effect size is inspected. Next, if the estimate is statistically significant, one is advised to infer a true effect and apply PEESE to estimate its magnitude. However, the statistical power of PET to detect an effect is unknown, and may be quite poor for sample sizes and effect sizes typical of psychology **?**. Given that a nonsignificant test result does not imply the truth of the null hypothesis, we are reluctant to privilege PET over PEESE. Thus, the present manuscript reports both PET and PEESE estimates for all meta-regressions. Readers are advised that if the null hypothesis is roughly true, PEESE will overestimate the true effect size, but that if the null hypothesis is false, PET will underestimate the true effect size.

The efficacy of PET-PEESE metaregression is supported by a simulation study by **?**,

This meta-regression technique has been previously applied by **?** to inspect the amount of evidence for "ego depletion," the phenomenon of fatigue in self-control. They found that after adjusting for small-study effects, PET-PEESE suggested an absence of evidence for the phenomenon. The authors therefore recommended a large-sample pre-registered replication effort, now supported by the American Psychological Society as the topic of the third Registered Replication Report

(http://www.psychologicalscience.org/index.php/publications/observer/obsonline/aps-announces-third-replication-project.html).

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

***p*-Curve.** Another novel technique for accounting for small-study effects is p-curve (Simonsohn, Simmons, & Nelson, 2014). p-curve estimates the true effect size by inspecting the distribution of significant p-values. When the null hypothesis is true (i.e. $\delta = 0$), the p-curve is flat: significant p-values are as likely to be between .00 and .01 as they are between .04 and .05. When the null hypothesis is false, the p-curve becomes right-skewed such that p-values between .00 and .01 are more common than are p-values between .04 and .05. The degree of right skew is proportionate to the power of studies to detect an effect, such that increasing sample sizes or larger true effect sizes will yield greater degrees of right skew. By considering the p-values and sample sizes of significant studies, p-curve can be used to generate a maximum-likelihood estimate of the true effect size.

One weakness of p-curve is that, in the presence of questionable research practices such as sequential data analysis, conditional use of covariates, motivated treatment of outliers, and moderator munging, an excess of p-values will gather close to the $p = .05$

threshold. This results in a flatter p-curve than would be found in more principled analysis, and thus p-curve will underestimate the true effect size in these circumstances. That aside, simulation work suggests that p-curve is quite effective at estimating true effect sizes [CITATION NEEDED].

Methods

We apply PET-PEESE meta-regression and p-curve effect size estimation to the C. A. Anderson et al. (2010) meta-analysis, using the meta-analytic data provided by those authors.[1] Because the data were analyzed using Comprehensive Meta-Analysis with the intent of testing for moderators, many studies were entered with separate rows for different outcomes or subsamples within studies. However, our current models assume that entire studies are censored or re-analysed and thus that each study should constitute a single observation. Thus, in the event that multiple effect sizes were entered for a particular study (e.g., effects on mean intensity and count of high intensity trials in the CRTT; separate simple effects for men and women), we aggregated these to form a single effect size for the study. For effects representing separate outcomes within a single sample, the outcomes were averaged. For effects representing separate subsamples within a study, the sample sizes were summed and a weighted average made of the subsample effect sizes. This parallels the behavior of the Comprehensive Meta-Analysis program used by C. A. Anderson et al. (2010). p-values were calculated via t-test, first dividing Fisher's Z scores by their standard errors to generate a t-value, then using that t-value to get a two-tailed p-value.

We then applied the meta-analytic adjustments. PET was performed by fitting a weighted-least-squares regression model predicting effect size as a linear function of the

[1] Since the publication of the C. A. Anderson et al. (2010) meta-analysis, a second meta-analysis has been published summarizing research published between 2009 and 2014 (Greitemeyer & Mügge, 2014). We had originally planned to include this meta-analysis in the present manuscript, but in the course of our research, found a number of errors. These authors are currently working to correct their meta-analysis, at which time we will apply these techniques to that research as well.

standard error with weights inversely proportional to the square of the standard error. Similarly, PEESE was also applied, predicting effect size as a quadratic function of the standard error and using similar weights. Finally, *p*-curve effect size estimates were generated using code provided by Simonsohn et al. (2014), entering a *t*-value and degrees of freedom parameter for each relevant study.

PET and PEESE estimates are provided regardless of whether statistically significant bias was observed according to recommendations by **?**, p. 20-21: "To be conservative, one should always use [the PET or PEESE estimate] even if there is insufficient evidence of publication selection because the Egger test [of publication bias] is known to have low power." Furthermore, simulations have suggested that the conditional application of meta-regression corrections (that is, applying them only when tests of bias attain statistical significance) tends to perform poorly compared to the unconditional application of such corrections, as a nonsignificant test result does not necessarily constitute firm evidence against publication bias (**?**). For similar reasons, we provide both PET and PEESE estimates regardless of the significance of the PET estimator. The power of PET to detect true effects seems questionable, and a nonsignificant PET result does not constitute strong evidence of no effect. The reader is encouraged to consider together the *p*-curve, PET, PEESE, and naive estimates in the context of the provided funnel plots and ongoing research into the efficacy of meta-analytic adjustments for bias.

Within the meta-regressions, all effect sizes were converted to Fischer's Z so as to fulfill the regression model's assumptions of normally-distributed effect sizes. Effect sizes are converted back to Pearson *r* for tables and discussion. All meta-regressions were performed using the 'metafor' package for **R** (**?**), using the `rma()` function to fit a variance-weighted model with an additive error term. *p*-curve estimates were similarly converted from Cohen's *d* to Pearson *r* for consistency of presentation.

Both *p*-curve and PET-PEESE are likely to perform poorly when there are few datapoints. Therefore, our analysis is restricted to effects and experimental paradigms with

at least ten independent effect sizes. Data and code have been made available online in the case that the reader nevertheless wants to generate estimates for more sparse datasets or explore the impact of our inclusion and exclusion decisions.

In addition to our analysis of the full dataset as provided by Anderson and colleagues, we perform leave-one-out sensitivity analyses, removing each datapoint one at a time and making all adjusted estimates. For each analysis, a supplementary tab-delimited spreadsheet is attached that lists the individual studies and the estimates when they are left out.[2]

Two studies were removed from the meta-analysis in all analyses. First, **?**, study 1 was removed because its entered effect sizes were unusually large for their precision (i.e., effects on aggressive behavior $r = .60$ and aggressive cognition $r = .53$), were highly influential on the meta-regression model, and most importantly could not be found as entered in the C. A. Anderson et al. (2010) dataset by inspection of the original article. Similarly, Panee and Ballard (2002) was removed because the study tested the effects of violent primes on in-game behaviors and not the effects of violent gameplay itself; therefore, it does not provide a relevant test of the hypothesis.

We reproduce estimates from C. A. Anderson et al. (2010) and apply *p*-curve effect size estimation and PET-PEESE metaregression to detect and adjust for small-study effects. Sufficient datapoints were available to re-analyze experimental studies of aggressive affect, aggressive behavior, aggressive cognition, and physiological arousal, as well as cross-sectional studies of aggressive affect, aggressive behavior, and aggressive cognition. Studies are further divided to create separate best-practices-only and all-studies estimates

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

per C. A. Anderson et al. (2010) as sample sizes permit.

Results

Results for all performed meta-regressions are summarized in Table **??**. Funnel plots with overlaid PET-PEESE regression lines and curves are provided in Figure **??**. We note that visual inspection of the funnel plot often reveals clear asymmetry, particularly in those subsets of studies that C. A. Anderson et al. (2010) selected as "best-practices" studies. Below, we discuss these statistics and describe the results of sensitivity analyses.

GUYS, THOUGHTS? IS IT WORTHWHILE TO REPEAT ALL THE STATS IN THE TEXT? I'M GETTING EXHAUSTED FROM COPYING STATS BACK AND FORTH AND I DOUBT ITS WORTH IT. HOW SHOULD I WRITE THIS SECTION? IS IT WORTH IT TO REPORT ALL THESE DIFFERENT ESTIMATORS?

Aggressive Affect

Experiments. Among studies selected as best-practices, *p*-curve estimated the true effect size as $r = .16$, substantially smaller than the original naive estimate of $r = .29$. Among the full sample of best- and not-best studies, the estimate was again $r = .16$. Leave-one-out sensitivity analyses are presented in supplementary table XXX.

PET estimated the effect as $r = -.12$ and $r = -.11$ in the best-practices and full samples, respectively. PEESE suggested $r = .14$ for best-practices, consistent with the *p*-curve estimate, but suggested only $r = .06$ for the full sample. Egger's regression test found significant small-study effects in both the best-practices and full samples, $p < .001$.

In sensitivity analysis, it became apparent that one study (**?**) had substantial influence over the meta-regression line, having an extremely large effect size estimate measured with modest precision. After removing this study, the small-study effects were still apparent (best practices, $p_{Egger} = .002$; all studies, $p_{Egger} < .001$), but meta-regression estimates rose such that PET estimated a more sensible null effect rather than a negative

effect (best-practices: PET $r = -.01$, PEESE $r = .17$; full sample: PET $r = -.05$, PEESE $r = .08$). p-curve was not influenced much by this exclusion, recommending $r = .13$ for best-practices and $r = .14$ for full sample. For the full spreadsheet of leave-one-out sensitivity analysis, consult supplementary file XXX.

Cross-sectional research. Insufficient studies were selected to conduct separate best-practices and full-sample analyses, so only the full sample was analyzed. p-curve estimated the effect as $r = .16$, slightly larger than the original naive full-sample raw estimate.

PET-PEESE suggested that effects were significant ($p < .001$), contaminated by small-study effects ($p_{Egger} = .049$), and slightly adjusted from the naive estimate (PET $r = .11$, PEESE $r = .14$).

In sensitivity analysis, it was found that several of the studies had substantial influence over the PET-PEESE model. The most influential of these was **?**; excluding this study caused the PET estimate to fall to nonsignificance and the effect size to be estimated as $r = .05$. Other influential observations (and the estimated effect size after their exclusion) included Matsuzaki, Watanabe, and Satou (2004, study 2, $r = .13$), and **?**, $r = .16$.

Aggressive Behavior

Experiments. Among studies selected as best-practices by C. A. Anderson et al. (2010), significant small-study effects were detected ($p_{Egger} = .007$), but a significant effect was not ($p = .126$). PET estimated the effect as $r = .081$, substantially smaller than that of the naive or trim-and-fill estimates. No studies were observed to have substantial influence on the effect size estimate; at most, exclusion of **?** raised the estimate to $r = .11$. p-curve similarly suggested that the effect was minimal, $r = .07$.

Among all studies, PET found a significant effect of violent games on aggressive behavior ($p = .001$) and no significant small-study effects ($p_{Egger} = .322$). PEESE

estimated the effect as $r = .157$, again smaller than that of the naive or trim-and-fill estimates. Again, no studies were found to be particularly influential, with r ranging from .15 to .17. p-curve again suggested a minimal effect, $r = .05$.

Cross-sectional research. Cross-sectional associations were more robust and less contaminated by small-study effects. Among studies selected as best-practices, p-curve suggested an effect size of $r = .27$, consistent with the original raw best-practices estimate. The PET estimate was significant ($p < .001$), and although small-study effects were detected ($p_{Egger} = .02$), PEESE recommended minimal adjustment ($r = .26$). Sensitivity analysis indicated that the estimate was largely robust to the inclusion or exclusion of single studies, with r remaining between .25 and .27.

Results were similar in the set of all cross-sectional studies. p-curve recommended $r = .23$. The PET estimate was again significant ($p < .001$), and PEESE recommended an only slightly-adjusted effect size ($r = .19$). Again, small-study effects were detected ($p_{Egger} < .001$). Sensitivity analysis suggested a number of influential studies, but even so, leave-one-out effect size estimates did not vary much, ranging from $r = .18$ to $r = .21$.

Aggressive Cognition

Experiments. Among experiments selected as best-practices, p-curve suggested minimal adjustments from the original naive meta-analysis, $r = .19$. On the other hand, PET found neither a significant effect of violent games ($p = .055$) nor a significant small-study effect ($p_{Egger} = .086$). The effect size estimate was $r = .11$, much smaller than the naive estimate of $r = .22$. Because the p-value is very close to the critical threshold, one might consider the PEESE-adjusted effect size estimate of $r = .18$. Because the effect was very near significance, sensitivity analysis suggested some rather variable estimates, as the removal of a single study could cause the p-value to cross the significance threshold. For example, exclusion of Bushman and Anderson (2009) caused PET to reach significance, leading to a PEESE estimate of $r = .19$. In the other direction, exclusion of C. Anderson

and Dill (2000) caused the effect size estimate to fall to $r = .06$.

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

Cross-sectional research. Among cross-sectional research selected as best-practices, *p*-curve again agreed with the naive estimate, $r = .19$. PET found a significant effect of violent games ($p = .001$) and significant small-study effects ($p_{Egger} = .013$). The PEESE-adjusted effect size estimate was $r = .15$, slightly smaller than the naive estimate. Sensitivity analysis detected some variability in effect size estimates. Exclusion of **?** caused the estimate to rise to $r = .17$, while exclusion of **?** caused the PEESE estimate to fall to $r = .13$. When C. A. Anderson et al. (2004) was excluded, the PET estimate fell sharply, no longer reaching statistical significance and recommending $r = .06$.

Among the full sample of cross-sectional studies, *p*-curve was consistent with the naive estimate, $r = .17$. PET again found significant associations with violent games ($p = .005$) and effects of sample size ($p_{Egger} < .001$). PEESE recommended an adjusted effect size of $r = .13$, moderately smaller than the naive estimate. Sensitivity analyses indicated two particularly influential observations: exclusion of **?** caused the estimate to rise to $r = .15$, whereas exclusion of **?** caused the PET estimate to no longer reach significance, yielding an estimated effect size of just $r = .04$.

Physiological Arousal

Experiments. Among the subset of best-practices experiments, PET detected neither an effect of violent games ($p = .227$) nor an effect of small studies ($p_{Egger} = .466$). PET recommended an adjusted effect size of $r = .13$, smaller than the $r = .18$ estimate

given by naive meta-analysis. Results in sensitivity analysis were quite variable, as might be expected of the small number of observations: estimates varied from $r = .08$ to $r = .27$. Perplexingly, p-curve suggested an effect larger than that of the naive meta-analysis, $r = .26$.

In the total sample of all experiments, PET did not detect an effect of violent games ($p = .942$) but did detect small-study effects ($p_{Egger} = .039$). The PET estimate of the effect size was $r = -.01$. Sensitivity analysis revealed minimal influence from individual studies, with the estimated effect ranging from $r = -.02$ to $r = .02$. Again, p-curve estimates were very different, suggesting an effect *larger* than that of naive meta-analysis, $r = .27$.

Unpublished dissertations

I'm going to move the dissertation analysis and funnel plots here.

Discussion

Our findings differ from those of C. A. Anderson et al. (2010) in two important ways. First, the original meta-analysis claimed that methodologically strong studies found larger effects than did methodologically weak studies. Instead, we find that best-practices studies yield estimates comparable to the full set of studies. Division of studies into best- and not-best-practices exacerbated funnel-plot asymmetry, leading to higher naive estimates but comparable adjusted estimates. Second, the original meta-analysis argued that there was evidence that the research findings were strong and not contaminated by bias. In our analysis, we find instead that the effect of violent video games on aggressive behavior in experiments is likely very small (r = .05–.10). That said, effects on aggressive affect and aggressive cognition in experimental and cross-sectional research seem stronger and more robust, although p-curve and PET-PEESE often disagree about the strength of the effect.

Currently, we believe that p-curve is the stronger meta-analytic technique. Although PET-PEESE is intuitive, easy to visualize, and draws upon more studies than just the

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

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

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

Limitations

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

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

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

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

Selection Bias in Meta-Analysis

We observe some instances of flexible application of the best-practices criteria offered by C. A. Anderson et al. (2010). Flexible application of the inclusion criteria may have

lead to preferential selection of studies with significant results. This selection bias could explain why the best-practices studies had larger naive effect-size estimates but comparable adjusted estimates.

p-curve estimates very similar effect sizes for both best-practices and all-studies samples. Recall that *p*-curve inspects only the studies that attained statistical significance. Inspection of the funnel plots reveals that the studies selected as best-practices are generally those studies attaining statistical significance; therefore, the studies considered by *p*-curve are mostly the same across the two samples.

Content validity. The first best-practices criterion is that the violent and nonviolent game must be sufficiently different in violent content. Application of this criterion was not consistent. In some cases, studies were excluded for having nonviolent games that contained very mild cartoon violence, while in others, nonviolent games containing substantial violence were included. For example, comparisons between the violent game *Mortal Kombat* and the nonviolent game *Sonic the Hedgehog* were discarded as not-best practices (e.g., **?**) because "the nonviolent game contained violence" (C. A. Anderson et al., 2010, supplementary materials). Another study comparing a racing game *Moto Racer* against the violent game *Tekken 2* (**?**) was excluded for similar reasons, but we were not able to find any violent content in *Moto Racer*. Meanwhile, other studies involving comparisons between violent and not-entirely-nonviolent games were included. Konijn, Nije Bijvank, and Bushman (2007) was included, although it used the game *Final Fantasy* as a nonviolent game. *Final Fantasy* appears to be as violent, or more violent, than *Sonic the Hedgehog*, so the simultaneous inclusion of this paradigm and exclusion of the *Sonic the Hedgehog* paradigm indicates inconsistency in the application of this criterion. Similarly, a study by **?** was included as best-practices despite comparing the violent *Grand Theft Auto 3* to the purportedly-nonviolent game *Simpsons Hit and Run*. While lighter in tone and less explicit than *Grand Theft Auto 3*, *Simpsons Hit and Run* nonetheless allows the player to punch other characters, steal cars, and run over pedestrians. This content

lead video game ratings boards to assign *Simpsons Hit and Run* a rating as appropriate for teens, not children. Thus, again, the application of this criterion seems much stricter for studies with significant results than for studies with nonsignificant results.

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

Measurement quality. Selection bias may also have been facilitated by the application of best-practices criterion 5: The outcome measure could reasonably be expected to be influenced by the independent variable if the hypothesis were true. For an example of selection bias, see C. A. Anderson et al. (2004, study 2). In this study, participants were assigned to play a violent or nonviolent game, then complete a competitive reaction-time task measure of aggressive behavior with either an ambiguously or unambiguously provoking confederate. A significant effect was found amount the 90 subjects assigned to the ambiguous provocation condition ($r = .25$), but not among the 90 subjects assigned to the unambiguous provocation condition ($r = -.03$). These 90 subjects with a nonsignificant effect were dropped from both the best-practices and not-best-practices meta-analyses.

When asked for comment, Anderson said "Only the ambiguous provocation condition was used because we now know that the unambiguous (increasing) provocation version of the task is not as sensitive to a variety of independent variables as is the ambiguous provocation pattern. In other words, the increasing provocation conditions don't meet Criterion 5." While it is possible that only one form of the task is sensitive to the manipulation, the meta-analysis does not seek to model such fine-grained moderators; at the least, the full sample should have been included in the full-sample meta-analysis. Furthermore, the validity or invalidity of measurements cannot be determined on whether

they provide the researcher with the desired $p < .05$ in an experiment. Finally, since a significant effect in either the ambiguous or unambiguous provocation group would be taken as evidence for an effect of violent video games, we are concerned that the selective exclusion of groups for not demonstrating such an effect risks introducing selection bias.

Selection bias may also influence which effect size among those reported was entered into analysis. As a general rule, it seems that C. A. Anderson et al. (2010) attempted to avoid subjectivity in effect size entry by averaging all reported effect sizes together. However, on several instances, effect sizes were not averaged together, but rather the single largest available effect size was selected. For example, in the aforementioned C. A. Anderson et al. (2004, study 2), the effect of violent games on the first trial of the CRTT was entered, but not the reported effect size on the other 24 trials of the CRTT. Selection of the largest effects risks capitalizing on chance and systematically overestimating the true effect.

Unfalsifiable predictions. We note further selection bias in the interpretation of violent games on physiological arousal. As presented by C. A. Anderson et al. (2010), violent games cause significant increases in physiological arousal, e.g. heart rate or blood pressure. However, in researching this meta-analysis, we became aware of studies in which null effects of violent games were excluded from meta-analysis. For example, in the best-practices studies by **?**, the violent and nonviolent versions of the game were not found to effect players' physiological arousal. Rather than present these findings as null results of violent games on physiological arousal, the authors presented this result as evidence that the violent and nonviolent games were matched stimuli. We observe a similar treatment in the meta-analysis: the null results on physiological arousal were omitted from the meta-analysis investigating effects of violent games on physiological arousal. We find this approach to be too flexible and forbids falsification of the theory, as concordant results are taken as evidence for the theory, but discordant results are excluded from consideration.

That said, although PET-PEESE estimates negligible effects on arousal relative to a

non-violent game, p-curve does estimate substantial effects. Because we suspect *p*-curve gives better estimates than PET-PEESE, we suppose that there are substantial effects of violent games on physiological arousal. Still, it would be helpul if it could be clarified when arousal is an inevitable consequence of violent games and when arousal is a confound that can be controlled.

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

Possible Data-Entry Errors

Some null findings were censored or miscoded. For example, in the course of the experiment reported in **?**, a nonexperimental assessment was also made of the effects of previous violent game exposure on aggressive outcomes. These nonexperimental effects were entered into the C. A. Anderson et al. (2010) meta-analysis, but were considerably changed from their report in the **?** manuscript. In the manuscript, nonsignificant effects of previous violent game exposure were reported for aggressive affect (study 1; $F(1, 66) = 0.78, r = -.11$), aggressive cognitions (study 2; $F(1, 57) = 0.02, r = .02$), and aggressive behavior (study 3; $F(1, 133) = 0.23, r = -.04$). However, as they appear in the meta-analysis, these were entered as two cross-sectional samples of violent game effects on aggressive behavior, this time with much larger effects than reported in the manuscript: $r = .33$ for studies 1 and 2 combined and $r = .23$ for study 3.

Unpublished Materials

The (C. A. Anderson et al., 2010) meta-analysis did make an attempt to collect and analyze unpublished studies (e.g., studies presented in dissertations or book chapters that did not undergo peer review). That the resulting analysis remained biased despite these

attempts gives us concern that searching for unpublished studies may not actually alleviate bias in meta-analysis.

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

Shortly after the publication of the C. A. Anderson et al. (2010) meta-analysis, there was some confusion as to the importance of unpublished research in meta-analysis. In a comment, **?** criticized the inclusion of unpublished research in the meta-analysis, arguing that such work is sometimes of dubious quality. These authors further criticized the purportedly-selective inclusion of *not yet published* research, such as articles under review or in press, and publications not peer reviewed, such as book chapters. In their reply, **?** describe unpublished studies as "studies not published in a peer-reviewed journal, although it could have been published in another outlet (e.g., book)." While they quote a passage from **?** stressing the importance of unpublished research as important to protection against confirmation bias and bias against the null hypothesis, the emphasis seemed to be nonetheless on book chapters and dissertations that were otherwise publicly available.

In our view, Drs. Ferguson and Bushman have both misinterpreted what is meant by "unpublished research." The unpublished research we are most often concerned about in meta-analysis are those studies that were conducted but never published in *any* form, whether journal article, dissertation, or book chapter. That is, we are concerned about "publication" in the most literal sense of *being made public.* Because studies that do not yield significant effects are less likely to be written, submitted, and accepted for publication, substantial parts of data may be missing from the scientific record. While

C. A. Anderson et al. (2010) report having searched thoroughly for unpublished materials, we note that the meta-analysis contains almost entirely studies that were published in at least one form or another (e.g., journal article, book chapter, or dissertation). Only two studies were found that were not published in any format. Given 20 years of research on a family of small effects, using small samples, it seems likely that there are more unpublished studies languishing in file drawers.

One particularly interesting publication format is the doctoral dissertation. Department requirements generally dictate that dissertations be submitted and published in a dissertation database regardless of whether or not that dissertation is later published as a peer-reviewed journal article. Dissertations, then, provide us with a sample of reported studies relatively uncontaminated by publication biases favoring significant results. We count XXX dissertations that did not later become journal articles. Effect sizes observed in XXX/YYY of these dissertations are nonsignificant. Given the number of dissertations that did not go on to be published in journals, one might wonder how many non-dissertation studies have been similarly conducted but not made public.

We note that few of these unpublished studies were accepted as best-practices research. Although we had hoped that the application of best-practices criteria would alleviate bias, recognizing well-performed research regardless of its results, it instead appears to have intensified bias.

Improving Research Quality

Historically, research practice has had a remarkable aversion to the null hypothesis. It wasn't until about 2011 that researchers realized the perils of publication bias and the value of the null hypothesis, reeling from the publication of Bem's (2011) evidence of precognition, the fraud of Diedrick Stapel, and insightful criticisms of the bias of typical research practices (Simmons et al., 2011; **?**; **?**; **?**). It is possible that researchers in this literature have been pressured by "obedient replication," the perception that those who

detect the effect are competent researchers and those that do not detect the effect are incompetent researchers. The pattern of nonsignificant findings among unpublished dissertations lends some credibility to this account. Dissertations failing to find a significant effect seem much less likely to have been published in journals or selected as best-practices studies.

A recent meta-analysis suggests that one particular researcher who fails to find effects obtains estimates significantly different from those of Anderson, Bushman, and disinterested third parties (Greitemeyer & Mügge, 2014). One might infer that this researcher cannot detect the effect because of incompetent or biased research, but another possibility is that this researcher is simply less averse to the publication of null results than others. That said, our own analyses suggest that this researcher's null results are sometimes underpowered and do not always provide much in the way of evidence for the null hypothesis (**?**). Again, transparent and collaborative research efforts could help to clarify whether the heterogeneity of effects across researchers is due to differences in methodology, competence, or bias.

Finally, we must register some skepticism of the more complex, interactive models of aggressive behavior that have been developed through this research literature. It seems that researchers quickly took the basic phenomenon for granted and began to cast about for more sophisticated models that would advance theory. In some cases, these more sophisticated models led to attempted conceptual, rather than direct, replications. It has been pointed out that the results conceptual replications can be difficult to appraise: if significance is attained, the replication is considered a success, but if significance is not attained, the replication may be considered invalidated by the changes in its paradigm.

In other cases, these more sophisticated models lead to the study of moderators and subgroups. When many moderators are tested, Type I error rates will rise substantially due to the problem of multiple comparisons. Post-hoc exploratory analyses of moderators are, of course, important and valuable (indeed, we have presented them ourselves in the

past), but become hazardous when presented as confirmatory or when patterns of statistical significance are taken to identify the validity or invalidity of the measures. We note that replication attempts of such interactions are exceedingly rare. Furthermore, if the main effect is as small as we estimate here, and if the moderating effects are on a similarly small scale, such tests of the interactions could be woefully underpowered, providing little positive predictive value and mostly generating Type I error.

The Competitive Reaction-Time Task

One popular measure of aggressive behavior, the Competitive Reaction-Time Task (CRTT), has been the topic of much discussion. While this measure is often used, it is rarely quantified the same way twice. It has been suggested that this variability in quantification is a form of p-hacking used by researchers to find larger, more significant effects (Elson, Mohseni, Breuer, Scharkow, & Quandt, 2014). Anderson and colleagues point out that studies using the CRTT find smaller, not larger, effects, suggesting that CRTT results are not inflated. [The following is my hypothesis and needs to be tested.] Our analysis finds that studies using the CRTT also feature larger sample sizes. It is possible, then, that the analysis and report of the CRTT is as biased or more biased than that of other measures, but that less bias is needed to reach statistical significance with these larger sample sizes. Thus, we maintain that there is clear need for validation of the noise-blast CRTT and for preregistration of the CRTT quantification that will be used in confirmatory research projects.

Implications for Theory

Even these adjusted estimates may still overestimate the true effect size due to the influence of confounds. Although it is often claimed that the observed effects are due to violent content alone (C. A. Anderson et al., 2004, e.g.,), the evidence for this claim is sometimes weak. For example, in many research projects, a small pilot study is conducted, finds nonsignificant differences in confounds, and concludes that $p > .05$ implies that the

null hypothesis is true. This method has been used to argue very different games are alike in all dimensions save violence; for example, the slow-paced puzzle game *Myst* was argued to be equivalent to the fast-paced shooter game *Wolfenstein 3D* (C. Anderson & Dill, 2000). Of course, $p > .05$ does not imply the truth of the null hypothesis, and a more appropriate Bayesian approach reveals minimal evidence of equivalence (**?**). Thus, small-sample pilot testing may not be sufficient to claim that the two conditions differ in violence alone. Application of confounds in analysis of covariance is a more promising approach, but this is also sometimes controversial (Miller & Chapman, 2001). When covariates are measured with error (e.g., with single-item Likert measures), substantial residual variance may be left behind and mistaken for variance associated with violence.

We have abstained from inspection of longitudinal studies as there are not enough data points to permit a good estimate. It is certainly possible, perhaps even likely, that there are detectable longitudinal effects of many hours of gameplay over time. This would seem more plausible than the prospect of a substantial and reliable effect obtained within fifteen to thirty minutes of gameplay. We echo the words of **?**, p. 51, "In many ways it is quite impressive that playing a violent video game for just 15-30 min, on a single occasion can have significant and measurable effects on aggressive behavior." Our tone, however, is different. Such an effect would be impressive, in the sense that it would be surprising and require substantial evidence to support. Our analysis suggests that the strength of evidence is not sufficient to support such a conclusion.

Researchers are encouraged establish the existence of the phenomenon before attempting to elaborate on the effect, its moderators, and its broader implications. If the effect is indeed as small as we estimate here (r = .16) then identifying meaningful and reliable moderators of the effect will be challenging. Moderators of the effect, if any, should be expected to have similarly small effects, and so may take impractically large samples to study. This may explain the counter-intuitive finding that effect sizes among children are not significantly larger than effects among adults: differences, if any, will be small, and are

likely to be obscured by research bias.

It may also be necessary to reevaluate our theories of aggressive behavior, specifically the General Aggression Model (GAM). According to GAM, aggressive feelings, thoughts, and behaviors are all closely related processes which feed into each other. Aggressive feelings are thought to inspire aggressive behavior, and aggressive thoughts increase the likelihood of aggressive behavior. The present analysis indicates that violent video games have minimal effect on aggressive feelings (a finding paralleled by Przybylski, Deci, Rigby, & Ryan, 2014) but that, depending on inclusion criteria and the exclusion of influential data points, there may be some effect on aggressive thoughts and some smaller effect on aggressive behavior. If this is the case, changes in aggressive behavior caused by violent game exposure would seem to be more related to aggressive-thought accessibility than to aggressive feelings. In our own research, we have attempted to collect measures of aggressive thoughts, feelings, and behavior from participants within a single session and found them to correlate poorly (**?**).

We echo the astonishment registered by **?**, p. 62: Given the theories and evidence in the rest of social psychology and media psychology, "Violent media can and must have some psychological impact on those who experience it, and probably does so via well-understood psychological processes."

The theories and evidence in the rest of social psychology and media psychology may be similarly weak. For example, the idea of "behavioral priming," e.g. that subliminally activating a thought influences automatic behavior (Bargh, Chen, & Burrows, 1996), holds a substantial position in the General Aggression Model [CITATION NEEDED]. Observing or participating in video game violence, it is argued, activates aggressive thoughts, which then cause increased aggression in behavior, particularly automatic behavior. Moreover, it is hypothesized that repeated exposure to violent media could cause aggressive thoughts to be chronically primed [CITATION NEEDED], a hypothetical extension of the phenomenon that is unique to this literature. However, recent attempts to replicate the phenomena

described by Bargh et al. have met with difficulty [CITATIONS NEEDED], and there is considerable skepticism about such direct priming effects in general. Nevertheless, proponents of violent game effects continue to cite Bargh's theory as support for video game effects without attention to the evidence against such mechanisms (Prot & Anderson, 2013; **?**).

Continuing the previous quote from **?**, p. 62, "Thus, for me, research in media violence no longer needs to establish whether such media can have a psychological and behavioral impact, but should instead rigorously examine the boundary conditions for such impacts." If the effects are indeed so small as we estimate, researchers will be hard-pressed to detect the boundary conditions. To detect r = .08 with 80% power, one-tailed, would require 960 subjects. To detect the small moderators that reduce the effect to insignificance may require a staggering amount of data.

References

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

Anderson, C., & Dill, K. E. (2000). Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. *Journal of Personality and Social Psychology*, *75*(4), 772-790. Retrieved from http://psycnet.apa.org/doi/10.1037/0022-3514.78.4.772

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

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

Bargh, J., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, *71*, 230-244.

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

Bushman, B. J., & Anderson, C. A. (2009). Comfortably numb: Desensitizing effects of violent media on helping others. *Psychological Science*, *20*(3), 273-277. Retrieved from doi: 10.1111/j.1467-9280.2009.02287.x

Egger, M. (1997). Bias in meta-analysis detected by a simple, graphical test. *British*

Medical Journal, *315*, 629-634. Retrieved from DOI: 10.1136/bmj.315.7109.629

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

Elson, M., Mohseni, M. R., Breuer, J., Scharkow, M., & Quandt, T. (2014). Press crtt to measure aggressive behavior: The unstandardized use of the competitive reaction time task in aggression research. *Psychological Assessment*, *26*(2), 419-432. Retrieved from 10.1037/a0035569

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

Ferguson, C. J., Rueda, S. M., Cruz, A. M., Ferguson, D. E., Fritz, S., & Smith, S. M. (2008). Violent video games and aggression: Causal relationship or byproduct of family violence and intrinsic violence motivation? *Criminal Justice and Behavior*, *35*(3), 311-332. Retrieved from 10.1177/0093854807311719

Greitemeyer, T. (2014). I am right, you are wrong: How biased assimilation increases the perceived gap between believers and skeptics of violent video game effects. *PLoS ONE*, *9*(4), e93440. Retrieved from DOI: 10.1371/journal.pone.0093440

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. Retrieved from 10.1177/0146167213520459

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

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

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

Nauroth, P., Gollwitzer, M., Bender, J., & Rothmund, T. (2014). Gamers against science: The case of the violent video games debate. *European Journal of Social Psychology*, *44*(2), 104-116. Retrieved from DOI: 10.1002/ejsp.1998

on Communications, C., & Media. (2009). From the american academy of pediatrics: Policy statement – media violence. *Pediatrics*, *124*(5), 1495-1503.

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

Prot, S., & Anderson, C. A. (2013). Research methods, design, and statistics in media psychology. In K. E. Dill (Ed.), (p. 109-136). Oxford University Press.

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

Sherry, J. L. (2001). The effects of violent video games on aggression. *Human Communication Research*, *27*(3), 409-431. Retrieved from 10.1111/j.1468-2958.2001.tb00787.x

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

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2014). Anchoring is not a false-positive: Maniadis, Tufano, and List's (2014) 'failure-to-replicate' is actually entirely consistent with the original. *SSRN*.

Valadez, J. J., & Ferguson, C. J. (2012). Just a game after all: Violent video game exposure and time spent playing effects on hostile feelings, depression, and visuospatial cognition. *Computers in Human Behavior*, *28*, 608-616. Retrieved from 10.1016/j.chb.2011.11.006