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

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

University of Missouri

Author Note

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

Abstract

Violent video games are theorized to be a significant cause of aggressive thoughts, feelings, and behaviors. A meta-analysis by Anderson and colleagues (2010) is thought to condense the research literature into robust and incontrovertible evidence that violent video games affect these outcomes in experimental, cross-sectional, and longitudinal research. In this meta-analysis, application of the trim-and-fill technique found minimal evidence of publication bias. However, there are now more sophisticated methods for the detection of, and adjustment for, publication bias. In the present manuscript, we examine previous meta-analytic evidence and apply these modern techniques for adjusting effect sizes in light of publication bias. Our conclusions differ from those of Anderson and colleagues in three salient ways. First, we detect significant publication bias in experimental research. Second, studies selected as being "methodologically stronger" do not find larger effects than other studies, but instead represent a subsample of the studies in which statistical significance was found. After adjusting for bias, there is no difference between the two estimates. Finally, 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 video games make their players more aggressive? Despite decades of research and hundreds of studies, the basic phenomena remain, at least for some, controversial. For some authors, what we term the *advocates*, the answer is definitively in the affirmative. For advocates, the effects are large, obvious, robust, and nearly ubiquitous. For others, what we term the *skeptics*, the research is not as clean nor as obvious as has been presented. Instead, skeptics point to a host of issues..... In the writings of the skeptics, the evidence for the violent video game effects is not as solid as claimed, in fact, it is paper thin.

Regrettably, the tenor of the debate has degraded recently as the advocates have en masse lost patience with the skeptics. Perhaps the most direct expression of this sentiment comes from **?** who writes:

> 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)

Here we see not only are the intentions of skeptics questioned, their scientific competence is too. Moreover, they are affiliated with fringe right-wing groups with a tenuous grasp of reality. Strasberger et al.'s quote is not isolated. **?** attempts to separate the advocates, termed "media-violence scientists" from the skeptics, who are less expert and whose conflict of interest keeps them from an honest assessment of evidence. **?** ignores the skeptics completely with a claim of consensus. Some have gone as far as to study video-game violence skepticism as a resistance-to-science akin to climate-change denial (see Greitemeyer, 2014; Nauroth, Gollwitzer, Bender, & Rothmund, 2014).

The advocates primary thesis is advanced by a meta-analysis from (C. A. Anderson et al., 2010). This meta-analysis covers xx studies spread across yy publication. The main findings are a large effect of video game violence on aggressive thoughts (r=, p<), a large effect of video game violence on aggressive feelings (), and a large effect of video game violence on aggressive behaviors. Moreover, these effects are found in a host of contexts including in experiments, in cross-sectional comparisons, and even in longitudinal research designs. **?** and Huesmann (2010, 2014) call the evidence in this corpus of studies as "decisive."

Despite this meta-analysis, there are still skeptics of causal effects of violent video games on aggressive outcomes. There are two class of critiques: one is about the *evidentiary value* of the results; the second is about the interpretability of the results. The evidentiary critiques are that the meta-analyses suffer from known difficulties including unaccounted publication biases and fortuitous selection criteria for inclusion. Added to this are concerns that the studies themselves suffer from questionable research practices including selective reporting of dependent variables and strategic inclusion of moderating covariates. The interpretability concerns are that violent video games differ from nonviolent games in more than violent content. Other hypothesized differences include arousal (Elson, Bruer, Van Looy, Kneer, & Quandt, 2013), competition (Adachi & Willoughby, 2011), and feelings of competence (Przybylski, Deci, Rigby, & Ryan, 2014). Therefore, it may be difficult to ascribe any effects to violent content rather than to these other differences. In correlational and longitudinal research, effects may reflect the possibility that aggressive individuals play more violent video games rather than the possibility that violent video game play makes individuals more aggressive. [NB: NOT SUCH A VALID CRITICISM OF LONGITUDINAL RESEARCH – WILLOUGHBY ET AL. LOOKED AT THE CROSS-LAGGED EFFECTS AND FOUND T1 VVG USE TO PREDICT T2 AGGRESSION RATHER THAN T1 AGGRESSION PREDICTING T2 VVG USE.]

In this paper, we focus solely on the evidentiary critiques, which are presented

subsequently in more detail. We ask whether there is solid evidence for the effect of violent video games on aggressive outcomes. We do not address the question of interpretation, and note that whether such effects are the causal outcome of the violent content or of some other factor is not known. Our approach is to reanalyze the meta-analysis of C. A. Anderson et al. (2010) with newer techniques that are more effective for assessing the degree of publication and selection bias. In their original meta-analysis, they applied the trim-and-fill procedure (**?**); in the present analysis, we apply meta-regression techniques (Egger, 1997; Stanley & Doucouliagos, 2014) and *p*-curve (Simonsohn, Simmons, & Nelson, 2014).

Given the nasty tenor of conversation between advocates and skeptics, we provide some background. The first two authors have extensive experience with video games and do not feel particularly aggressive. They came to graduate school to study social psychology, and at the start of their research believed the video-game violence effect is plausible but not certain. The last author, a cognitive psychologist, would like to know whether he should let his twelve-year-old daughter buy the Grand Theft Auto game she covets. He relented though more out of trust for her judgment than out of any deep skepticism of the literature. We are motivated by curiosity, not ideology. We are happy to support or refute video-game violence from the reported meta-analysis as the case may be.

Data were provided upon request by the corresponding author of C. A. Anderson et al. (2010). We use their original dataset without modification of the raw values or moderator codings. The purpose of the present study was not to question the competence of the original meta-analytic work (which, indeed, appears to have been done well), but rather to glean additional insight through re-analysis of the original dataset and through interpretation by a third party. As ever in meta-analysis, the crucial value is not trust, but rather transparency (CITE STEPHEN SENN, 2015). Furthermore, because novel meta-analytic techniques for handling publication bias are continually developed, it is beneficial to reanalyze past meta-analyses with novel statistical techniques (**?**).

The Evidentiary Critiques

Researchers performing meta-analysis must always deal with thorny statistical issues that may affect the conclusions. The issues here are *publication bias* and the *selection of studies*. We deal with these in turn:

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 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 (**??**). One might infer 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.

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

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

The efficacy of PET-PEESE metaregression is supported by a simulation study by **?**, who find that Peters meta-regression has less bias than naive random-effects meta-analysis or trim-and-fill. However, the [ETC ETC ETC]

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 (i.e., Peters metaregression, CITATION NEEDED). In the current report, we use as our effect size estimate Fisher's Z with standard error $\frac{1}{\sqrt{N-3}}$, consistent with the original analysis of Anderson and colleagues. Because this standard error is not a function of the effect size, we avoid the problem of an inherent relationship between precision and effect size that might otherwise contaminate the metaregression.

***p*-Curve.** Another novel technique for accounting for small-study effects is p-curve (Simonsohn et al., 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, an excess of p-values will gather close to the $p = .05$ threshold. This results in a flatter p-curve than would be found if studies had been reported without p-hacking, and thus p-curve will underestimate the true effect size in these circumstances. That aside, simulation work suggests that p-curve is quite effective at estimating true effect sizes [CITATION NEEDED].

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

Unpublished Materials

The above techniques describe statistical techniques for inspecting publication bias in research – ways to look for the influence of unpublished research and its influence on the estimate. We now describe the importance of unpublished research and a ovel way to estimate its prevalence.

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, **?** described unpublished studies as "studies not published in a peer-reviewed journal, although it could have been published in another outlet (e.g., book)." Although they quoted a passage from **?** stressing the importance of unpublished research as important to protection against bias, 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 parcels 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. In our analyses, we highlight unpublished dissertations and how they fared in meeting best-practices criteria.

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] The original authors' separation of studies by study design (experimental, cross-sectional, longitudinal) and by study outcome (affect, behavior, cognition, arousal) is sensible and accurate, and we maintain these distinctions in our re-analysis.

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 of publication bias

[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.

assume that entire studies are censored or re-analysed per their statistical significance; thus, each study should constitute a single observation. Thus, in the event that multiple effect sizes were entered for a particular study (e.g., effects on mean intensity and count of high intensity trials in the CRTT; separate simple effects for men and women), we aggregated these to form a single effect size for the study. For effects representing separate outcomes within a single sample, the outcomes were averaged. For effects representing separate subsamples within a study, the sample sizes were summed and a weighted average made of the subsample effect sizes. This parallels the behavior of the Comprehensive Meta-Analysis software used in the original analysis. p-values were calculated via t-test, first dividing Fisher's Z scores by their standard errors to generate a t-value, then using that t-value to get a two-tailed p-value.

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

[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.

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

Results

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

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

"best-practices" studies. Below, we discuss these statistics and describe the results of sensitivity analyses.

Egger's regression test

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

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

Adjusted effect sizes

Results of the p-curve and PET-PEESE analyses are supplied in Table XXX alongside naive fixed-effects and random-effects meta-analytic effect size estimates. Again, our in-progress simulation work suggests that p-curve may be the least biased and most efficient of these estimators. However, a weighted combination of several estimators often outperforms any single estimator. Therefore, we suggest that the reader consider all five estimates and apply her own weights in deciding for herself what seems the most likely true effect in each subsample.

Contrary to the conclusions of the original authors' naive estimates, p-curve does not think that best-practices studies measure a larger true effect than do not-best-practices

studies. In all cases save one, best-practices and not-best-practices studies received similar adjusted estimates; in the case of correlational studies of aggressive behavior, best-practices studies were estimated as measuring a slightly larger effect.

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

There is one notable case in which p-curve and PET-PEESE seem to agree on the estimate. When inspecting effects on aggressive behavior in experiments, both techniques estimated that the true effects were very small and likely not meaningfully different from zero. Notably, these estimates are highly consistent with some recent reports by the new generation of violent-media researchers (Przybylski et al., 2014; **?**).

Sensitivity analysis

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

Aggressive Affect: Experiments. Among experiments of aggressive affect, it was apparent that one study (**?**) had substantial influence over the meta-regression line, having an extremely large effect size estimate measured with modest precision. After removing this study, the small-study effects were still apparent (best practices, $p_{Egger} = .002$; all studies, $p_{Egger} < .001$), but meta-regression estimates rose such that PET estimated a more sensible null effect rather than a negative effect (best-practices: PET $r = -.01$, PEESE $r = .17$; full sample: PET $r = -.05$, PEESE $r = .08$). p-curve was not influenced much by this exclusion, recommending $r = .13$ for best-practices and $r = .14$ for full sample.

Aggressive Affect: Correlational. Among cross-sectional studies of aggressive affect, it was found that several of the studies had substantial influence over the PET-PEESE model. The most influential of these was **?**; excluding this study caused the

PET estimate to fall to nonsignificance and the effect size to be estimated as $r = .05$. Other influential observations (and the estimated effect size after their exclusion) included Matsuzaki, Watanabe, and Satou (2004, study 2, $r = .13$), and **?**, $r = .16$.

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

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

Aggressive cognition: Experiments. [THIS NEEDS TO BE REWRITTEN BECAUSE I'M TRYING TO GET AWAY FROM THE NHST IN PET-PEESE.] Because the effect was very near significance, sensitivity analysis suggested some rather variable estimates, as the removal of a single study could cause the p-value to cross the significance threshold. For example, exclusion of 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.

Aggressive Behavior: Correlational. Among best-practices cross-sectional studies of aggressive cognition, exclusion of **?** caused the estimate to rise to $r = .17$, while exclusion of **?** caused the PEESE estimate to fall to $r = .13$. When C. A. Anderson et al.

(2004) was excluded, the PET estimate fell sharply, no longer reaching statistical significance and recommending $r = .06$. In the full sample, sensitivity analyses indicated two particularly influential observations: exclusion of **?** caused the estimate to rise to $r = .15$, whereas exclusion of **?** caused the PET estimate to no longer reach significance, yielding an estimated effect size of just $r = .04$.

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

Unpublished dissertations

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

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

Chi-square tests were highly significant for all tests. The relationship between statistical significance and publication status was highly significant such that unpublished dissertations were much less likely to have found statistical significance than published studies (liberal test, $p = 3.94 \times 10^{-6}$; conservative test, $p = 4.02 \times 10^{-6}$). Similarly, the relationship between publication status and best-practices inclusion was highly significant such that unpublished dissertations were far less likely to be included as best-practices than

published studies (liberal test, $p = 2.17 \times 10^{-8}$; conservative test, $p = .002$). Although we had hoped that the application of best-practices criteria would alleviate bias, recognizing well-performed research regardless of its results, it instead appears to have intensified bias.

Discussion

Our findings differ from those of 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*. (At worst, the player can bump into another driver in such a way that both drivers fall off their bikes; neither driver is injured, and the player suffers a time penalty.)

Meanwhile, other studies involving comparisons between violent and not-entirely-nonviolent games were included. 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 to favor the inclusion of significant results and the exclusion of nonsignificant results.

Flexibility in the application of this criterion may have contributed to selection biases, inflating the naive meta-analytic estimate relative to the adjusted estimate. A better approach might be to have manipulations rated by research assistants naive to hypotheses or to study results, or to seek a statistical quantification of the difference in

violence between games, such as a Cohen's d describing a manipulation check.

Measurement quality. Selection bias may also have been facilitated by the application of best-practices criterion 5: The outcome measure could reasonably be expected to be influenced by the independent variable if the hypothesis were true. For an example of selection bias, see 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. Returning again to C. A. Anderson et al. (2004, study 2), the effect of violent games on the first trial of the CRTT was entered (mean difference = 1.07), but not the reported effect size on the other 24 trials of the CRTT (trials 2-9, mean difference = 0.08; trials 10-17, mean difference = 0.04; trials 18-25, mean difference = 0.19). Again, Anderson and colleagues may think that this first-trial-only measure is the most appropriate measurement, at least for this particular study. We are less certain. Selection of the largest effects risks capitalizing on chance and systematically overestimating the true effect. There may be some flexibility involved in the decision to select one trial from a set of twenty-five, to be reported in only one half of the total sample. As Elson, Mohseni, Breuer, Scharkow, and Quandt (2014) point out, not every study uses 1st-trial-only CRTT behavior as the outcome; perhaps the decision to use this particular outcome is contingent on its statistical significance.

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.

Omissions

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

Unpublished Dissertations

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 (sometimes deliberately), 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.

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

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 et al., 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, perhaps not in this manuscript.] 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 stimulus 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. Thus, insofar as effects remain after adjustment for small-study effects, they may still be contaminated to some degree by confounds.

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.

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

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

of data.

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. A similar phenomenon may explain why, contrary to what intuition would predict, interactive media do not seem to have greater effects than noninteractive media [CITATION NEEDED].

In sum, the research literature as analyzed by (C. A. Anderson et al., 2010) seems to contain greater publication bias than their trim-and-fill analyses and conclusions would indicate. This is especially true of those studies which were selected as using best practices, as the application of best-practices criteria seemed to be influenced sometimes by the results of the study. Effects in experiments seem to be overestimated, particularly those of violent video game effects on aggressive behavior. Rather than accept these estimates as the "true" effect sizes, we recommend instead a preregistered collaborative research effort and prospective meta-analysis. In this research effort, preregistration and collaboration will both be indispensable. In the absence of preregistration and collaboration, the two well-defined camps of proponents and skeptics may each find results that support their conclusions and refuse to believe the results of the other camp. If we are to "advance the debate" over violent game effects, we must do it not by silencing the other party, but by getting each party to sit down together, design an experiment, and say in writing for all to see, "I agree that this is the appropriate research design. My theory predicts that the result shall be this; his theory predicts that the result shall be that. Together, let us see who is right, and move on."

References

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

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

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

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

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

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*.

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