BUL-2015-0509  
Overestimated Effects of Violent Games on Aggressive Outcomes in Anderson et al. (2010)  
*Psychological Bulletin*  
  
Dear Dr. Hilgard,  
  
Thank you for submitting "Overestimated Effects of Violent Games on Aggressive Outcomes in Anderson et al. (2010)" for review and consideration for publication in *Psychological Bulletin*. The editorial board has completed its review. In addition to reading the manuscript myself, I was extremely fortunate to have received reviews from outstanding experts in the field both in terms of substantive dimensions and methodology. Their feedback is unusually detailed and constructive, and I am grateful to them for their time and service to the field.   
  
All reviewers found something to like about your work, although they ranged in their recommendations from “reject” to “minor revisions” and each argues that needed is more evidence or else improved procedures to lead to clearer evidence and clearer conclusions. At present, it seems none of the reviewers is completely convinced that Anderson et al. (2010) overestimated the impact of violent games on aggression. Based on my independent reading, my own view is closer to “Major Revisions,” and this letter provides my rationale as well as sums up what I see as the larger problems the reviewers identified, though of course each review should be consulted in its entirety.  
  
Assumptions of the Database  
  
There are important concerns about effect size calculations (see especially Reviewer 3’s remarks about your pp. 15 and 16). Until such problems are addressed, the substantive conclusions and contributions of your work will not be clear. Admittedly, much of what Reviewer 3 has raised about effect size calculations has not entered conventional meta-analytic practice. Given that you took Anderson et al.’s database at face value, I realize that this task makes the project larger than you had intended at the outset, but it seems important to examine these assumptions carefully. If nothing else, taking a random, sizable sampling of reports in the database would permit you to see if these assumptions matter. To the extent that the assumptions do not matter, then they need not take a central role in your revision. In this event, put the relevant details in an online Appendix. (Given that many outcomes in the literature were dichotomous, it might be wise, as well, to consult McGrath and Meyer, 2006, who show that effect size metric often matters and that interpretations of effect size magnitude may differ depending on the choices taken.)  
  
“Publication Bias” Estimators and Heterogeneity  
  
Once assumptions in effect size calculations are cleared up, then other matters may be addressed and these are at the center of your current re-analysis. Even given the current state of the database, Reviewers 2, 3, and 4 are adamant that heterogeneity took a less prominent place in your interpretations than it should; I strongly concur.   
  
Lau et al. (2006) have made perhaps the most compelling qualitative illustration of this problem, which is that what looks like publication bias may result from substantive differences among studies. There also may be confounding between publication bias and moderator levels. Such assumptions are not controlled in the lion’s share of “publication bias” statistics. Indeed, nearly all of these statistics might well be better labeled “asymmetry” statistics, as they assume a single population of effect sizes (Johnson & Eagly, 2014; Johnson, Low, & LaCroix, 2013). It is for reasons such as these that “publication bias” statistics are used mainly as sensitivity analyses (Reviewer 4), not as ways to divine what values missing studies might take or that such studies truly are missing. Reviewer 4 also argued that, “in order to attribute the asymmetry to publication bias and related biases, it is crucial to exclude the possibility that the small study effects are due to germane methodological and substantive reasons.” This concern helps flesh out, in different words, my own concerns about interpreting publication bias statistics in the face of heterogeneity. As Reviewer 4 argues, small samples may well differ methodologically from large samples.  
  
In the face of such concerns, Reviewer 4’s recommendation to move to a strategy like Sterne and Egger's (2005) seems quite sensible (Weinhandl & Duval, 2012, offer an alternative, though I am not sure whether a software solution is readily available). The caveat, of course, is that Anderson et al.'s, 2010, results did not have much evidence of moderation, despite their use of fixed-effects assumptions, which are statistically over-powered. [Apropos, Reviewer 3’s recommendation to use random-effects assumptions also seems prudent because you have heterogeneity in almost every sub-group of effect sizes, although Anderson et al., 2010, p. 158) reported that doing so “yielded very similar results.” (Note that statistical power is low to detect heterogeneity when k of studies is small.)]  
  
Neither Anderson et al. (2010) nor your meta-analysis reported I^2 values related to the mean effect sizes (cf. Higgins & Thompson, 2002; Huedo-Medina et al., 2006). Indeed, your report omits heterogeneity statistics entirely. I^2 values correct for the fact that the value of Q depends in part on k of studies, and therefore are useful to judge whether there is more heterogeneity in one sub-group of effect sizes compared to one or more others. Indeed, I^2 can even replace Q, as long as a 95% CI for I^2 is given. (Reporting tau^2 is reasonable, but it is more difficult to interpret.) I recommend, like Reviewer 4, where relevant, that you add I^2 and its 95% CI, and interpret results accordingly. I am less of a fan of interpreting the magnitude of I^2 with benchmarks. Although zero implies homogeneity and values higher than zero imply that effects vary more than sampling error would predict, it is arbitrary to label 25 "small", 50 "medium", and 75 "high" heterogeneity. (Indeed, 100 is not possible, except with rounding up, because it is equivalent to infinity.) Yet, you *can* say that one sub-group of effects has greater heterogeneity than another, such that effects in that sub-group are more variable due to unknown causes (assuming no significant moderation). In essence, some studies find the predicted effect and others do not. Some may even reverse the predicted effect.  
  
Reviewer 1 noted “in the area of violent video game research it is definitely possible to report non-significant findings” in the published literature. I urge you to think about the problem through that lens: To the extent that non-significant findings appear in the published literature, is it not evidence against publication bias? At least it means that publication bias should be less of a concern, or perhaps it is not the over-riding concern. The fact that there were so many unpublished dissertations may not be so much publication bias as it is authors not attempting to publish their findings (e.g., because their jobs post-graduation do not expect scholarly achievements). Here Reviewer 4’s suggestion to move to contour-enhanced funnel plots (for funnel-plot Figures that you retain in your manuscript) is quite sensible (as you did in your Figure 7; you can use different symbols for published and unpublished studies). Directly comparing unpublished with published studies is the most direct test of publication bias. I found it curious that such a comparison did not appear in the Anderson et al., 2010, meta-analysis. I’ll add a small quibble: On page 5 you say, “publication bias, is the phenomenon that studies with statistically significant (i.e., p < .05)”; yes, but strictly speaking, p<.05 is a conventional but not necessary criterion, scholars often deviate from this criterion, such as with the Bonferroni correction. In sum, “(e.g., p < .05)” would be more appropriate.  
  
p-Hacking and p-Curves   
  
As Reviewer 2 asserts, “These are relatively new techniques that have yet to be fully scrutinized by the meta-analytic methods community.” The lesson is to interpret these statistics with appropriate caution (and to think about more homogeneous subsets of the data). The same concerns about use of “publication bias” test may well be at issue in “p-hacking” and “p-curves.” That is, it is possible that deviations from linear p-curves result from heterogeneous sub-populations of effect sizes. It would seem important to solve the problems noted in the prior section about handling heterogeneity before interpreting p-curve findings.  
  
[Side notes: (a) I agree with Reviewer 3 that “p-hacking” is a sub-optimal term, though to some extent we are stuck with convention. (b) You say (e.g., p. 13), “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”, yet, strictly speaking, a p-value of exactly .00 is not defined. I believe you are rounding p-values that are less than .005 to “.00”; scholars might often do so in practice, but it is a misleading convention. (c) Finally, I strongly suggest omitting Figures 1 and 2, as your paper should focus on substantial matters rather than on educating about a method.]  
  
Methodological Quality  
  
Consistent with Reviewer 3’s comment about your page 9: Be more careful in terms of thinking about methodological quality, both in terms of the originating studies and in meta-analyses of these studies. Anderson et al. did address study quality by using a several-item scale; Valentine (2009) provides an incisive review about the methodological quality of studies and how best to use this information in meta-analysis. As you will see there, it is not clear how most methodological problems will affect the magnitude of effect sizes. (See also Johnson, Low, & MacDonald, 2015, who build on that strategy, analytically, and who agree with Valentine about the possibility that even one methodological flaw in a study may invalidate or bias it.) More generally, there are also standards for quality in meta-analysis of growing use, perhaps especially the AMSTAR scale (Shea et al., 2007).   
  
Transparency and Data Sharing  
  
It is consistent with APA ethical codes for authors to provide data within a reasonable time frame after publication, as Anderson et al. did in the present case. I agree with the reviewers that we must promote transparency in our science, but it is too early in the process to ask, as Reviewer 4 suggested, that Anderson and colleagues provide an online archive of their database and other materials. I will defer this action for now.  
  
Rhetoric and Tone  
  
In making your revisions, I ask that you address the matters the Reviewers and I have raised with as much care and politesse as possible. Certainly, the reviews and your re-analysis make clear that there may be problems in Anderson and colleagues’ methods (and possibly their conclusions). I realize that to some extent you have followed prominent conventions in conducting your re-analysis of Anderson et al.’s (2010) database, so what I am advocating here is to help steer convention toward better accuracy in meta-analytic methods and interpretations thereof. Still, as the introduction in your current manuscript already implies, we have learned from and can benefit more from Anderson and colleagues’ work. (Reviewer 3 makes a similar point about your p. 4; indeed, statements like “Good work deserves re-analysis; decisive work requires re-analysis” ought to be jettisoned entirely.) Finally, it would of course be unfair to “blame” Anderson and colleagues for using sub-optimal methods when many of these methods were not well known at the time or have emerged into usage sense then. Instead, they made what appears to be a good faith effort, with the largest yet database of evidence, to address a question of fundamental importance to societies around the world. Ideally, what I would want to see in your work is a contribution that builds upon the earlier work rather than tears it down. Your estimates of the mean effect size may well end up being somewhat smaller than the original effect, but does that mean that the effect is not real or not meaningful?   
  
Conclusion  
  
Please let me know within a week if you will be submitting a revision that addresses the points I've outlined. If you chose to resubmit, please do so within 3 months of receipt of this letter.  
  
When submitting the revision, if you decide to do so, please do so electronically at [Psychological Bulletin](http://bul.edmgr.com/) and log in as an author. You should see a menu item entitled "Submissions Needing Revision," where you will find your submission record there under BUL-2015-0509.   
  
If you resubmit, be sure to indicate in your cover letter that your work is a resubmission. Please include with your revision an item-by-item description of the changes that you made to the manuscript in response to each of the comments. Please detail where changes are made and quote the manuscript as appropriate.  
  
I look forward to hearing from you. Please let me know if you have any questions.  
  
Sincerely yours,  
  
Blair T. Johnson, Associate Editor  
Psychological Bulletin  
  
  
References  
  
Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. Statistics in Medicine, 21(11), 1539-1558.  
  
Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I² index? Psychological Methods, 11(2), 193-206.  
  
Johnson, B. T., & Eagly, A. H. (2014). Meta-analysis of social-personality psychological research. In H. T. Reis & C. M. Judd (Eds.), Handbook of research methods in social and personality psychology (2nd Ed., pp. 675-707). London: Cambridge University Press.  
  
Johnson, B. T., Low, R. E., & LaCroix, J. M. (2013). Systematic Reviews to Support Evidence-Based Medicine (2nd Edition) by Khalid Khan, Regina Kunz, Jos Kleijnen and Gerd Antes [Book review]. Research Synthesis Methods, 4(1), 102-108. (doi:10.1002/jrsm.1071)  
  
Johnson, B. T., Low, R. E., & MacDonald, H. V. (2015). Panning for the gold in health research: Incorporating studies’ methodological quality in meta-analysis. Psychology & Health, 30(1), 135-152.  
  
Lau, J., Ioannidis, J. P., Terrin, N., Schmid, C. H., & Olkin, I. (2006). Evidence based medicine: The case of the misleading funnel plot. BMJ: British Medical Journal, 333(7568), 597–600.  
  
McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: the case of r and d. Psychological methods, 11(4), 386-401.  
  
Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., ... & Bouter, L. M. (2007). Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC medical research methodology, 7(1), 10.  
  
Valentine, J. C. (2009). Judging the quality of primary research. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), Handbook of research synthesis and meta-analysis (2nd ed., pp. 129–146). New York, NY: Russell Sage Foundation.  
  
Weinhandl, E. D., & Duval, S. (2012). Generalization of trim and fill for application in meta‐regression. Research Synthesis Methods, 3(1), 51-67.  
  
=================  
  
Reviewer #1: The authors argue that a previous meta-analysis summarizing the effects of violent video game play has considerably overestimated the impact of playing violent video games on aggressive behavior, cognition, and affect. This is a well-written manuscript that presents some interesting findings. I am also sure that the authors are correct in that publication bias (and other biases) should lead to an overestimation of the effect. As with any field in (social) psychology, it could be proven difficult to publish a finding that a relation between variables was not significant so I'm sure many authors abstained from writing up a study when violent game play was not related to increased aggression (although in the area of violent video game research it is definitely possible to report non-significant findings). I am not sure, however, whether I should share the authors' view that the impact of playing violent video games is negligible.  
  
For one, media exposure has been demonstrated to influence the consumers (e.g. Valkenburg et al., in press, Annual Review of Psychology) so it is unclear why people should be susceptible to all kinds of media exposure, with the exception of video game violence. Moreover, all theories of human aggression propose that exposure to violent video games is associated with increased aggression so there is no theoretical reason for the claim that there is no relation between violent video game play and aggression.  
  
Looking at the empirical basis, arguably the relation between violent video game exposure and aggressive behavior is the most interesting one. Whereas the p-curve analysis suggests that the experimental effect is small, the PEESE estimate for both best practices and the full sample is .15, which is not so far away from the Anderson estimate. It is also important that the estimates for the correlational studies are similar to the Anderson estimates. (It's a shame that the longitudinal data could not be re-analyzed.) Because the estimates are so different, more evidence is needed (in my view) before coming to a final conclusion.  
  
Overall, I feel that the authors' conclusion that effects are minimal is not warranted. As noted above, I am sure there are some biases in previous violent video game research, but I am not convinced that these biases are so extreme that there is no relation between violent video game exposure and aggression at all.  
  
=================  
  
Reviewer #2: This manuscript reanalyzed Anderson et al.'s (2010) meta-analysis of the effects of violent video games on aggression. The focus of the reanalysis was to assess the role of publication selection bias in the reported estimates using relatively new methods. These methods detect substantially larger publication selection bias than what was found by Anderson et al., significantly reducing the size of the estimated effect of violent video games. Overall, I feel that this manuscript makes an important contribution to this area of research but think that the work can be strengthened in several ways as detailed below.  
  
The introduction claims that PET, PEESE, and p-curve "provide for better adjustments" for publication selection bias. While this may be true, it is far from certain. These are relatively new techniques that have yet to be fully scrutinized by the meta-analytic methods community. Furthermore, they have the distinct disadvantage of not making use of all of the available evidence.  
  
The discussion of publication bias addresses many important aspects of this issue but seems to miss the issue of outcome (dependent variable) selection bias. It does get mentioned but only in a single sentence at the end of a paragraph on p-hacking. The role of selective outcome reporting in driving publication bias should be more clearly highlighted.  
  
The discussion of the weaknesses of the trim-and-fill method misses an important issue: study heterogeneity. The trim-and-fill methods is most effective in homogeneous or at least only mildly heterogeneous distributions. The method becomes less effective as the effect size distribution becomes more heterogeneous. This issue is mentioned in discussing funnel plots but the idea is not adequately developed.  
  
In arguing against the trim-and-fill method, the authors state that "studies are not likely to be censored on the basis of the effect size, but rather, on the basis of their statistical significance". The implication is that the trim-and-fill method is simply based on effect sizes. This is incorrect. The trim-and-fill method is examining the relationship between the effect size and its standard error and as such is focusing on statistical significance, at least indirectly. A complication is that this effect size and its significance may not be in agreement with the statistical test performed by the original author. However, the p-values used by the methods employed in this study are based on the calculated effect sizes and standard errors, not the p-values report by the original authors (see top of page 16). Thus, this argument against the trim-and-fill applies as well to the methods used in this manuscript.  
  
The discussion of existing methods of estimating and adjusting for publication selection bias fails to discuss the method developed by Vevea and Hedges (1995, Psychometrika). This method has conceptual similarities to the methods used in this article but estimates the adjusted effect size using all available data. Comparing the methods used in this article with the results from the Vevea and Hedges method would enhance the value of this manuscript. Furthermore, this method deserves a discussion in the introduction.  
  
On page 16, the section on adjusted estimates does not clearly articulate whether a fixed or random effects model was implemented. The first paragraph of this section implies a fixed effects model whereas the second implies a random effects model. A random effects model seem more plausible for these data. Furthermore, the authors should state what method of estimating tau^2 was used (assuming a random effects model is use) and this should be reported in the tables.  
  
=================  
  
Reviewer #3: I have read Anderson et al. (2010) and the present article carefully and my impression is that the authors of this re-analysis raise some notable concerns about the original findings. Most importantly, contrary to the conclusions by Anderson et al. (2010), there does seem to be some credible evidence for "small-study effects". This in itself does not necessary indicate the presence of publication bias and/or p-hacking [1]. For example, as noted by Anderson et al. (2010) and the present authors, if the size of studies is chosen according to properly executed power analyses, we would in fact expect to see an inverse relationship between outcomes and sample sizes (and so if authors engage in the recommended practice of planning a study to achieve sufficient power, we are actually building small-study effects into our literature!). However, the finding of small-study effects in these data does make it at least more plausible that certain selection mechanisms may have introduced bias into the results (at least more plausible than if no evidence for small-study effects was found).  
  
[1] Personally, I am not too crazy about the term, for various reasons. For one, the term "data dredging" has been around for a long time and is also more encompassing, since not all practices of data dredging are aimed at influencing the p-value. This aside, the term is used in a derogatory manner, while "hacking" is not something done for malicious purposes (despite how the term is often used, or rather, misunderstood by the popular media). But okay, this may just reflect my own personal biases.  
  
So, while I think raising attention to this issue is important, I have a number of comments about the paper, some of which I think will deserve some attention by the authors.  
  
Abstract: "In contrast, the cross-sectional literature appears relatively robust and unbiased." I am having some difficulties matching up this statement with the results. Figures 3-7 and Table 2 all seem to indicate the presence of small study effects, also for cross-sectional studies. Table 1 is a bit more mixed, in part due to the number of different methods. This aside, I assume the authors are referring to "publication bias" here when they talk about "unbiased". But there are other biases that could be present and some of those biases (e.g., confounding) are more likely to be present in cross-sectional studies than experimental and longitudinal research.  
  
p. 4: "In many ways, it would be remarkable if violent video game effects were not at least somewhat overestimated, as biases that overestimate effect sizes are common in science." Given the contentious nature of the topic at hand, I looked carefully for statements in the manuscript that could be interpreted as rhetoric. For the most part, I think the authors did a nice job of avoiding rhetorical arguments, but appeal to what would (or would not) be remarkable is irrelevant here. Let me quote from Huesmann (2010) a sentence that I find equally objectionable: "It requires a tortuous logic to believe that children and adolescents are affected by what they observe in their living room, through the front window of their house, in their classroom, in their neighborhood, and among their peers but are not affected by what they observe in movies, on television, or in the video games they play." This is the same rhetorical device.  
  
p. 5: "Publication bias ... is an especially dangerous problem for meta-analysis [...]". As far as I am concerned, it is no more and no less a problem for meta-analysis as any other method for reviewing the literature (including reading a single article). It may be more difficult to demonstrate how publication bias can lead to wrong conclusions in a narrative review, but the underlying problem is the same and just as severe.  
  
p. 6: "Second, the authors found 16 dissertations which had yielded nonsigificant [sic!] results and subsequently gone unpublished, but only one unpublished non-dissertation study. Given that dissertations likely represent a minority of all studies conducted on violent games [...]". Can the authors provide any evidence that this is the case? And is it really necessary to make this claim? I think the finding reported can stand on its own.  
  
p. 9: "Funnel plots are without doubt the most important graphical summary of the quality of a meta-analysis." Funnel plots are useful (but not so much when there is heterogeneity or the number of studies is small), but I do not think it is the most important graphical summary of a meta-analysis or of its quality (and by the way, how would one even define the quality of a meta-analysis?).  
  
p. 9: "In the absence of small-study effects or heterogeneity [...]". Replace the "or" with "and" (both need to be absent).  
  
p. 12: "In PET, a weighted linear regression is fit to describe the relationship between effect size and precision, much like the Egger regression test." Not just "much like", but this is exactly what is done in the Egger regression test. The only difference is that the Egger regression test is focused on the relationship between the effect sizes and their precision (i.e., the slope of the regression model), while PET then looks at the intercept as an estimate of the "true effect".  
  
p. 15: "We calculated t-values by dividing values of Fisher's z by their standard errors [...]". That is not the proper way to compute a t-statistic based on a correlation coefficient. Instead, one should compute t = r \* sqrt(n-2) / sqrt(1-r^2), which follows a t-distribution with n-2 degrees of freedom under H0: rho = 0. I am not sure though if all of the values used are really Pearson product-moment correlations. Anderson also included partial correlation coefficients, for which this equation would not apply (and again, dividing the corresponding Fisher's z value by the SE does not yield a t-value then either). As an approximate procedure (with appeal to asymptotics), the Fisher's z value divided by the standard error could be treated as following a standard normal distribution (under H0).  
  
p. 16: "[...] all effect sizes were converted to Fisher's z [...]". A couple points:  
  
- Is this what Anderson et al. (2010) did in their analyses? I cannot find anything in that article that indicates whether raw correlations or Fisher's z values were used in the meta-analysis. So, a clarification (whether this is the same as was done in the original meta-analysis or not) would be useful.  
  
- How did the authors compute the SEs for partial correlation coefficients transformed with the Fisher transformation? Note that, based on Aloe (2014), and in analogy to the Fisher transformation for Pearson product-moment correlations, it should be 1/sqrt(n-p-3), where p is the number of predictors (but assuming that p is small relative to n, using the usual formula of 1/sqrt(n-3) would give a similar answer).  
  
Aloe, A. M. (2014). An empirical investigation of partial effect sizes in meta-analysis of correlational data. Journal of General Psychology, 141(1), 47-64.  
  
- Many of the correlations in the dataset are actually based on two-group comparisons, so t-values (for the independent samples t-test) and/or standardized mean differences must have been transformed to correlations. However, those are not Pearson product-moment correlations, but point-biserial correlations. The Fisher transformation is not a proper variance stabilizing transformation for point-biserial correlations and the SE is not 1/sqrt(n-3) then. In fact, based on Pustejovsky (2014), the asymptotic variance of a z-transformed point-biserial correlation can be shown to be equal to 1/(d^2 + w^2) \* ((n1 + n2) / (n1 \* n2) + d^2/(2\*(n1+n2))), where d is the standardized mean difference, n1 and n2 are the group sizes, and w^2 = (n1 + n2)^2 / (n1 \* n2).  
  
Pustejovsky, J. E. (2014). Converting from d to r to z when the design uses extreme groups, dichotomization, or experimental control. Psychological Methods, 19(1), 92-112.  
  
- In fact, Pearson product-moment correlations and point-biserial correlations are not estimating the same underlying type of parameter and therefore it is inherently illogical to combine them (one would not average a log odds ratio and correlation coefficient either). I do not know whether these two types of coefficients are being mixed here and/or by Anderson et al. (2010), but if they are, then don't. Instead, if one regards the grouping variable as a dichotomized continuous variable, then one can transform the point-biserial correlation into a biserial correlation coefficient, which does estimate the same parameter as a Pearson product-moment correlation. Again, see Pustejovsky (2014).  
  
p. 16: "P-curve estimates were similarly converted from Cohen's d to Pearson r for consistency of presentation." I am confused now. How does Cohen's d come into play here? And how were the values transformed?  
  
p. 19: "[...] than did Anderson et al.'s trim-and-fill esitmators". Check spelling.  
  
p. 25: "We suspect that this asymmetry is caused in part by ambiguities in study design and stimulus selection that make it impossible to report null results on aggressive affect." But Figure 3 shows that there are studies reporting null results for aggressive affect.  
  
Figure 3: In the left-hand side plots, I see one point equal to a Fisher's z value of about 1.33, so back-transformed this would be a correlation of about .87 (or since these are experimental studies, this may be a point-biserial correlation). Is this not an outlier? And is this the Ballard and Wiest (1996) study the authors discuss on page 20? Also the rightmost point in the plot at the top right seems kind of extreme (also in Figure 5, but less so).  
  
Table 3: What is "Mixed" statistical significance? And why does this not apply to the other 3 tables?  
  
=================  
  
Reviewer #4: In this submission the authors reanalyze the data from a highly cited meta-analysis published in Psych.Bull. in 2010 that uncovered a mean effect of playing violent video games on aggressive cognition and aggressive behaviors. Given my expertise in publication bias and the detection/correction thereof and my relative lack of knowledge on the substantive issues related to the effects of playing violent video games, I will focus my review on the statistical issues related to potential publication bias and related biases.   
  
Taken together, I side with the authors that there are clear indications of exaggerated effect sizes in Anderson et al. possibly due to publication bias and related biases. The text is well written, the results are clearly described, and the most of the relevant issues are discussed in an open and rigorous manner. So I feel that the current submission has merit and should be published to correct the overly strong conclusions made by Anderson et al. Nonetheless, there are several issues that need to be addressed.   
  
The current state of knowledge is that trim and the fill procedure used by Anderson et al. is a weak method that quite often misses important indications of publication bias and related biases (note that the literature review on this aspect is limited because there are many others besides Simonsohn et al. who have made this point). A widely accepted sign of publication bias is funnel plot asymmetry. And such asymmetry is strikingly clear in the funnel plots based on Anderson et al.'s data. As the current authors rightly indicate, funnel plot asymmetry (small study effects) can occur for methodological and substantive reasons other than publication bias. One way of dealing with this is to select subsets that are fairly homogeneous in terms of methodology, as the current authors did, for instance by selecting only experiments of a particular type. Nonetheless, in order to attribute the asymmetry to publication bias and related biases, it is crucial to exclude the possibility that the small study effects are due to germane methodological and substantive reasons. The traditional example of substantive reasons for asymmetry/small study effects is one in which a treatment is more effective for patients that are more severely affected by a disease. Such more severely affected are also typically more difficult to sample and may even be quite rare, leading to small study effects. Other methodological reasons may relate to the set-up of the experiments with more intense treatments being much more labor intensive, leading not only to smaller samples (under scarce resources) but also to larger effects in the studies with smaller sample sizes. The authors are aware of these issue, but I would have liked to read more discussion of this aspect as it is crucial to the PET-PEESE method and the Egger test. Note that it is crucial for a meta-analyst to specify such potentially relevant moderators a priori (which unfortunately is uncommon in meta-analyses that seldom use pre-registered hypotheses and analyses) on the basis of theory and methodological considerations. Someone who claims that funnel plot asymmetry is due to methodological and substantive factors needs to make a strong case and has to show that relevant study-level characteristics can predict study outcomes. This is one of the advantages of using meta-regressions to explain funnel plot asymmetry (see my point below), because it enables one to see whether funnel plot asymmetry (i.e., the prediction of effect sizes by studies' standard error) remains after controlling for these moderators (substantive/methodological). If such a case cannot be made, and the methodological moderators cannot explain the asymmetry, it is completely reasonable to conclude that the asymmetry is due to publication bias. So this issue warrants more discussion and perhaps additional analyses of subsets or meta-regressions that include the SEs as predictors besides other potentially relevant moderators that are associated with the studies' sample sizes.  
  
The authors used the Egger test. Egger himself has now abandoned this version of the test in favor of the weighted regression as described by Sterne & Egger (2005) in the famous publication bias book. I would have chosen the weighted version as it is more in line with the PET-PEESE approach. As mentioned, the Sterne and Egger approach has the additional advantage that it allows the inclusion of substantive and methodological moderators in an effort to try to explain asymmetry due to such more germane sources,  
  
In my view, the funnel plots in the current submission should be replaced by contour-enhanced funnel plots that depict the 95% confidence band around the null hypothesis. The 95% band is now centered around the estimated mean effect sizes, but this is not very informative because the estimated effect is possibly inflated because of publication bias. With contour enhanced funnel plots, the reader can more directly see that the dots (study outcomes) mostly lie just above the p=.05 threshold. Hence, publication bias related to significance should be more apparent in such contour-enhanced plots.  
  
On a related note, it would be valuable to add Ioannidis & Trikalinos' (2007) exploratory test for excess significance to the analyses. This analysis has been widely used and studied by Greg Francis in the last three years. Although it is not without problems, it certainly has some diagnostic value and so would be interesting to see whether it highlights an excess of significant results in the violent game literature as reviewed by Anderson et al.  
  
Van Assen et al. (2015, Psych Methods) have developed the p-uniform method that is similar to the P-curve method by Simonsohn et al. (it was based on the same rationale), but has some advantages above the p-curve method. Specifically, the p-uniform routine (see <https://rvanaert.shinyapps.io/p-uniform>) functions a bit better than the p-curve estimator in some cases where there are many p-values close to p=.05. More importantly, p-uniform offers 95% confidence intervals around the estimated mean effect and it provides another test of publication bias (both of which are unavailable in p-curve). So I suggest the authors to also run the analysis with p-uniform to get more indications of the severity of publication bias and to add a sensitivity analysis to the results already presented. Note however that both p-curve and p-uniform are based on the notion of a fixed effect and so will not function well when there is substantial heterogeneity. For that reason it would be worthwhile to also add indications of heterogeneity (Q test results, estimates of tau, or Higgins I^2) to the results of the traditional fixed or random effect models.   
  
I was glad to see that the authors discussed the limitations of the methods they used. Methods to detect and correct for publication bias invariably have strong assumptions and so are often seen mostly as sensitivity tools. Therefore, it is valuable to add some additional analyses to the current set of results. If funnel plot asymmetry (Sterne & Egger test), PET-PEESE, the test for excess significance (Ioannidis & Trikalinos, 2007), p-curve, and p-uniform (van Assen et al., 2015) all point to publication bias, the conclusion would be hard to avoid that the results presented in Anderson et al. (2010) are inflated and that we need large-sample pre-registered studies to settle the question whether playing violent video games have an effect on aggressive cognitions and aggressive behaviors.   
  
The current submission presents crucial results of a reanalysis that rightly correct some widely disseminated claims made by Anderson et al. The reanalysis highlights that data from meta-analysis should always be included in the article itself or made available to the readership for scrutiny in a data repository. Unfortunately, the data were not shared in this submission or in Anderson et al. but were supposedly only available upon request from Dr. Anderson. This is a suboptimal practice that should be banned from academic publishing because it disallows others (including myself as reviewer and interested reader) to double check the analysis that lies at the heart of the contention. Nowadays, it is very easy to share data as appendix or online (in a repository like OSF), and so I urge the editor to politely ask Dr. Anderson to share the data of his widely cited paper in the interest of scientific debate and openness.