BUL-2015-0509R1

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 once again received reviews from the same outstanding experts in the field. Their feedback is constructive and often quite detailed (one even re-did a simulation you reported!), and I am grateful to them for their time and service to the field.

Despite considerable work on your team’s part in revising your paper, and despite eliminating several threats to your conclusions with admirable simulation work, there is still fairly marked disagreement in the reviewers’ advice about your manuscript, with Reviewer 1 recommending “Reject” and the other 3 either “Accept” or “Minor revisions”. My own view is somewhere between “Minor revisions” and “Major revisions,” so I must decline this opportunity to publish this particular version of your paper. Clearly, you need to clean up the issues that Reviewers have noted, so please read each carefully and respond accordingly, but let me sum up the main issues the reviewers raised, and add my own views:

1. How best to interpret your Results? Reviewer #1 re-stated your conclusion as “violent video game exposure does not have an impact on aggressive behavior,” but such a definitive statement does not seem to occur in your manuscript, whereas the Anderson et al. (2010) meta-analysis definitively claimed that it (viz., “the effect”) exists. Instead, I see you making more ambiguous claims, that “the effect” is overestimated in the Anderson et al. results, but not that the effects do not exist. As you document, small experimental studies are more likely to do so than larger experimental studies. Moreover, some of your bias-detection statistics leave a statistically significant effect (e.g., PEESE for experiments, aggressive behavior; PEESE for the best experiments, aggressive affect; importantly, these effects are homogeneous, with I2=0). Thus, at least some of your results suggest that in fact “the effect” does exist, albeit at a smaller magnitude than Anderson et al. obtained. Finally, the cross-sectional analyses seem largely in accord with the original meta-analysis. Thus, there is little (any?) need to claim that the cross-sectional effects are “over-estimated,” which the title and running title of your paper suggest is the case regarding the Anderson et al. analysis. Clearing up these issues would enhance your paper in terms of interpreting your data (which the Reviewers found, in the current version, only moderate). Thus, you should resolve these ambiguities throughout the paper, though of course do so after resolving interrelated issues, such as the next point.

Okay. We make it super clear that it’s just the experiments that are bullshit. The correlations are a different kind of bullshit, being incapable of inferring causality.

As far as the idea that we’ve found evidence of the effect: PET and PEESE each begs the question by assuming that there isn’t a true effect (PET) or that there is a true effect (PEESE). In the case of aggressive behavior, p-curve agrees with PET, suggesting that there’s no (significant) effect, r < .10. I’m not sure the degree to which significance testing is a good idea, given that the sensitivity and specificity of PET and PEESE are unknown and may not match their nominal Type I error rates. SO WHAT THE FUCK.

2. Assumptions in the database analyzed. None of the Reviewers focused on this issue, but I want you to check the extent to which removing studies from the database affects your conclusions. You have a database with three fewer studies and at least four fewer effect sizes; three of the ¬effects were rather large (r=.60, .53, and .57). A critic might well fix on the fact that your database differs as the reason for your more conservative results about “the effect”: First, it could well be the omission of these effects that makes your bias-corrected mean effects smaller (in addition to bias-correction). Second, it may be that their negligible trim-and-fill results depended on the presence of these effects, and your analysis does not attempt to replicate those findings. If your analysis showed that trim-and-fill is still missing the bias, now we have reason to suspect that differences in the database are not responsible for the different results. You may well find other ways to investigate these issues (e.g., sensitivity analyses). Finally, why is it not possible to re-calculate the correct effect size for the Graybill, Kirsch, and Esselman (1985) study (instead of the manipulation check wrongly included in the Anderson et al. database) and include it in your results? A critic could argue that you are showing your own selection bias by eliminating it, assuming that this study still found a non-trivial effect size.

We’ve created a supplementary table conducting naïve fixed-effects meta-analyses and trim-and-fill analyses as Anderson et al. did. Notably, their trim-and-fill analyses were restricted to the “best experiments” and “best partials” datasets.

Only a few of the effect sizes we removed were part of the dataset subjected to trim-and-fill. Panee and Ballard (2002) is treated as a best-practices study of aggressive affect and physiological arousal, but a not-best-practices study of aggressive behavior. Graybill et al. is listed as not-best-practices, and so is not considered in their trim-and-fill analyses. Matsuzaki et al. is a cross-sectional study, not an experiment.

Thus, the dataset for our publication analyses regarding experiments differs by only one effect: Panee and Ballard (2002)’s effect on aggressive affect. Again, we point out that this study is not hypothesis relevant because the manipulation is an aggressive prime and the outcome is violent game play, rather than violent game play being the manipulation.

Anderson et al. report trim-and-fill results for best experiments *r+* = .294, with zero imputed studies. Excluding Panee & Ballard, we get naïve FE result *r* = .289, and trim-and-fill imputes 6 studies to the left, reducing the effect size to *r+* = .238. Including Panee & Ballard, we get naïve FE result *r* = .308, a little bit higher than their naïve FE result. Whereas their trim-and-fill imputed zero studies, however, ours imputes 6, reducing the effect size to *r+* = .247. We’re not sure what’s responsible for these differences. Considering how nakedly asymmetrical the funnel plot of best-practices aggressive affect experiment effect sizes are, it seems odd that trim-and-fill would have imputed zero studies. Perhaps there was an error on their part?

We double-checked the Graybill, Kirsch, and Esselman (1985) study. The main outcomes were categorizations of the direction and type of participants’ aggression in the Rosenzweig Picture-Frustration Task. That is, all participants were placed into one category of each; the total amount of aggressive thought was not measured. This leads us to think this experiment is not a relevant hypothesis test. Anderson and colleagues entered the non-manipulation-check outcomes as having a net zero effect size (r = -.02, r = .02), so we are being conservative with respect to our hypotheses by excluding it.3. Square pegs in round holes? Although the Reviewers found your manuscript at least moderate in terms of its appropriateness for Psychological Bulletin, they also found its contribution of new knowledge relatively modest, with most rating it 2 or 3 (out of 5; one gave a higher score, but I think that judgment is probably based on the statistical dimensions). It is true that the current paper reads more like one from the days when this journal still had a methods section (which perhaps makes it a “square peg in a round hole”), and I suspect that is one reason Reviewer 1 is as opposed to publication as s/he is, pointing out that all extant theories predict that stimuli aggressive beget behaviors aggressive. As this reviewer implies, readers of the Bulletin are likely to “expect more” than they see in this revision. Accordingly, I ask your team to expand your Discussion about future directions. I am an outsider to this literature, but some comment about past theories’ predictions would seem in order. I am not asking for a dramatic expansion, but, I noticed that you have done your own primary-level work in this domain, and you have, after all, taken a new statistical look at the literature. (Your introduction can be more forward looking in raising the reasons why extant theories posit that "the effect" should exist.)

Sure. Previous theories argued that brief gameplay sessions “activated aggression-related networks” which would later have unconscious and automatic influences on behavior, i.e. social priming. Sometime around the third or fourth refuted Bargh study the field stopped mentioning this.

Personally, I think that social priming provides an excellent theoretical framework for thinking about these effects, but not in the way one might expect. It’s not “social priming” in that minor manipulations have big unconscious effects; no, it’s “social priming” in that experiments get tortured until a significant p-value is obtained. It’s “general strain theory”or whatever O’Boyle et al. called it that made people torture their dissertations into publishable shape.

The Anderson et al. meta-analysis essentially rested on these theories’ prediction of a main effect for “the effect” and had little in terms of potential moderators. Yet their analyses and yours also documented heterogeneity in effects (at least in some types of research in this domain): What is driving these discrepancies? In other literatures, interventions tend to have more dramatic impact on members of vulnerable populations than on those who buffer such effects by their own resources (e.g., empathy) and the resources of others connected to them (e.g., imagine parents scolding bad behavior). Similarly, Reviewer #3 suggests that more research with rigorous methods (large samples and preregistrations) is needed to understand the effects of violent video games on aggression, but such efforts should be cognizant of unexplained heterogeneity in “the effect.” Similarly, very recent research has shown that instabilities in replication findings may often be attributed to hidden moderators (see Van Bavel et al., 2016, PNAS, 10.1073/pnas.1521897113).

[Honestly, buddy? Who the fuck knows. Considering that all these studies are run on college kids, I’m guessing the heterogeneity is methodological, not personological. There’s clear methods heterogeneity.]

The Van Bavel paper is controversial (see X, X, X, CITATION NEEDED).

The issue in this literature is not that some experiments do find the effect and others do not. The issue is that the studies selected as best-practices by Anderson et al. almost always *do* replicate (i.e., find statistical significance), even when they should not (i.e., very poor power). There’s no heterogeneity in effects on behavior in experiments once you’ve accounted for small-study effects. For affect, some of the heterogeneity is caused by the Ballard & Weist outlier.

Finally, although no reviewer raised this issue thus far in the review process, it seems germane: I think a critic may find it a weakness that you did not also analyze the covariate-adjusted cross-sectional effects that Anderson et al. also examined. I agree with your logic not to analyze them as you did the other findings, but it seems to me that at a minimum you should discuss your findings in relation to Anderson et al.'s finding that controlling for covariates reduced "the effect" substantially. Your analyses suggest that there are no small-study effects among the cross-sectional studies, but it would seem even those effects are also over-estimating "the effect." And these adjusted factors are relevant to a theory of when "the effect" is bigger or smaller. (Another way forward would be encouraging pooling of individual-level databases in individual-participant meta-analysis, which could examine such factors with the greatest sensitivity and statistical power.)

[Sure. The partial effect sizes are small, and that’s interesting. See Ferguson’s paper, too, although I think it’s much too sloppy w/r/t what’s being adjusted for in the partials. ]

We now mention individual-participant meta-analysis as another benefit of open data archival.

4. Other comments. I am attaching a copy of your manuscript with further suggestions. Please contact me directly if you have any questions.

In short, 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 by July 6th; if you need more time, please let me know. When submitting the revision, if you decide to do so, please do so electronically at Psychological Bulletin 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-2015-0509R1.

If you resubmit, be sure to indicate in your cover letter that this 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,

Blair T. Johnson, Associate Editor

Psychological Bulletin

Reviewers' comments:

Reviewer #1: The revision is clearly improved. In particular, I had an issue with how the authors present their findings (a bit they know it all) and this is toned down. Nevertheless, I still don't think this paper should be published in Psychological Bulletin. It is an informative manuscript, yes, but for a publication in Psychological Bulletin, I would expect more.

The main message is that Anderson et al. underestimated the impact of publication bias. Although I believe that this is likely, I'm wondering if there is any (social) psychological literature that does not suffer from some degree of publication bias. I'm quite sure that one could reanalyze all previous (social) psychological meta-analyses and make the same point.

We think that it is very likely that many (social) psychological literatures suffer from some degree of publication bias. That is why there is presently such an interest in bias-correction techniques for meta-analysis and preregistered replication studies. These things help us to determine how bad the publication bias is, and whether or not an effect is likely to exist. With regard to the present study, we feel that the literature on violent game effects should aspire to reflect the truth, not simply to be only as bad as the rest of social psychology.

The second message is that the impact of violent video game exposure on aggressive behavior is negligible. Yes, the data seem to suggest that, but still this finding is hard for me to believe. For example, according to weapons effect, even a picture of a weapon increases aggressive behavior. But playing video games where serious (virtual) violent acts are committed should have no impact. This is difficult to imagine. Likewise, simple primes, such as the color black or exposure to heat words, increase aggression. Moreover, different kinds of media exposure influence the consumers, so I'm still wondering why people are susceptible to all kinds of media exposure, with the exception of video game violence. All in all I doubt the authors' conclusion that violent video game exposure does not have an impact on aggressive behavior.

I am having trouble finding citations for the effects Reviewer 1 mentions. I cannot find papers claiming that seeing heat-related words increases aggressive behavior. I know it has been claimed that priming by heat words causes increases in the *perception* of aggressive behavior (DeWall & Bushman, 2009). However, new evidence suggests that this finding is not replicable (McCarthy, 2014). Other research suggests that uncomfortable temperatures can cause aggression (e.g., Anderson, 1989), but there is clearly a difference between actually being hot vs. simply seeing hot words. I’m also having some trouble finding claims that the color black causes aggression. There are correlational studies suggesting that teams wearing black or red jerseys receive more penalties (Frank & Gilovich, 1988), but follow-up studies contest this claim (Caldwell & Burger, 2010).

I am skeptical also of the weapons priming effect. Some results suggest flexibility in analysis. For example, Bartholow, Anderson, Carnagey, & Benjamin (2005) uses a very eccentric quantification of aggressive behavior in the CRTT, possibly a form of significance chasing (again, see Elson et al., 2014). Others use subgroup analyses when significant main effects could not be found. For example, Turner et al. (1975) find effects of weapon primes on driving behavior, but only for a very particular complex contrast within one level of another manipulated factor (Study 1), or after splitting subjects by gender and by model of car (!, Study 2). Apparently Drs. Benjamin and Bushman are preparing a meta-analysis on the weapons priming effect. I look forward to seeing if there’s anything to it beyond publication bias.

We now explain that we suspect the short-term effects of violent games to be much like the short-term effects of other aggression primes, in that they may be null effects overestimated through significance-chasing.

As I noted in my previous review, almost all theories of human aggression suggest that exposure to violent video games is associated with increased aggression. Of course, these theories may be wrong (as the authors comment in their letter to the editor), but I would expect a theoretical treatment why there should be no link between violent game exposure and aggression (this is Psychological Bulletin!), and the authors remain completely mute in this regard.

We now explain our Null Theory, which argues that behavior is too complex and unpredictable to be reliably influenced by minimal manipulations such as media priming. We refer also to general strain theory (Merton, 1938; O’Boyle, Banks, Gonzalez-Mule, 2014) and/or obligatory replication (Ionannidis, CITATION NEEDED), as an explanation for the potential significance-chasing these studies represent. We’re not going to go so far as to claim that, say, Social Learning does not happen or that the General Aggression Model is false. It seems reasonable to say that people learn behaviors and reward-contingencies by watching others, or that people who feel more aggressive and think more aggressively are likely to act more aggressively. Rather, we think that brief exposure to violent games is not sufficient to influence these mechanisms.

In general, we do not think that there needs to be any particular theory to explain why many things are simply not effective. No theory is needed to explain *why* acupuncture is not an effective treatment for cancer – it simply isn’t.

===============

Reviewer #2: This version of the manuscript is greatly improved and addresses my concerns.

The wisdom and beauty of Reviewer 2 are without parallel.

===============

Reviewer #3: "1. Reviewer 3 suggests that the funnel-plot asymmetry could be the consequence of the combination of an unbiased research literature, heterogeneous effect sizes, and a priori power analysis. This is true, but unlikely in the present analysis for several reasons."

First, I would like to clarify that I did not mean to imply that this is the more likely explanation for finding funnel-plot asymmetry in the present context or that it could explain away the present findings. It is but one of many possible explanations whose plausibility one can (and should) debate. At the very least, it is an interesting thought experiment to consider, because it would imply that funnel-plot asymmetry is an inherent feature of the scientific literature. This aside, I think the authors' responses are reasonable, except for: "First, we understand the use of power analysis to be quite uncommon in 2010 and earlier." The first edition of Jacob Cohen's highly influential book on power analysis (Statistical Power Analysis for the Behavioral Sciences) was published in 1969 and sample size calculations have been done for decades, even if they are not reported in articles. Hence, "Power analyses are rare in the studies synthesized by Anderson and colleagues" (p. 29) is not convincing as far as I am concerned. But I would tend to agree that there are other/more plausible explanations for the funnel-plot asymmetry in the present context.

[If people were doing power analyses, then you wouldn’t see studies with N = 40 like Bartholow & Anderson (2002).]

"11. Reviewer 3 suggests that we have misestimated the variance of Fisher's Z. See IC above."

I read about the results of the simulation study with interest. Unfortunately, I was not able to find the supplementary file (neither as part of the submission nor on the OSF repository), so I ended up rewriting the simulation to double-check these results. Results were similar, but did not match exactly what the authors show in their figure (within an appropriate margin of error). For example, for n=40, delta=0.5, and an allocation ratio of 0.75, the empirical standard deviation of the d-to-r(point-biserial)-to-z transformed values was 0.159 based on 10^6 simulated values, but the value shown in the figure is slightly above 0.18 (which is too far off even when considering the simulation error). No error is involved in computing 1/sqrt(40-3) =~ 0.164, while the equation of Pustejovsky depends on the observed value of d, so we can take the average of the estimated standard errors, which yields ~0.155 in this scenario, which in turn appears very close to what is shown in the figure. At any rate, it is indeed the case that both equations provide similar results under the simulated scenarios. In fact, it isn't all that surprising that 1/(n-3) works as the variance estimate for small effects, since the equation in Pustejovsky simplifies to 1/n when d=0 (regardless of the allocation ratio). On the other hand, things start to break down for the 1/(n-3) equation for larger effects (while the Pustejovsky equation still yields consistent estimates), but one could argue that this is not so relevant in the present context, since the underlying true effects are unlikely to be so large for this to become a real issue.

I understand the authors' desire to use the 1/(n-3) equation, since it eliminates the inherent relationship between the r-to-z transformed values and the corresponding variances when using the equation from Pustejovsky. However, it needs to be emphasized that 1/(n-3) is purely an ad-hoc approximation that should \*not\* be used in general, while the equation given by Pustejovsky has been properly derived based on valid statistical principles. These types of ad-hoc approximations irk me quite a bit, since I have seen too many examples where Fisher's r-to-z transformation was applied to correlations that are \*not\* Pearson product-moment correlations between two continuous variables, yet the resulting values were treated as having variance 1/(n-3). This is in fact wrong, even when it may work in some cases as an adequate approximation. Unfortunately, numerous books on meta-analysis (including Borenstein et al., 2009; and the CMA software) perpetuate these types of inaccuracies, which is a real shame.

Obviously, this paper is not the place to open this can of worms, since it distracts from what this paper is actually about. Also, in the present context, I think the approximation will be good enough for government work (not sure if one wants to be associated with that though). However, I think it would be important to at least add a footnote on page 18 (where the 1/sqrt(n-3) equation is introduced) to explain that this equation is only an approximation for the standard error of d-to-r-to-z transformed values that is technically correct only if the true effect is zero, but that still works adequately as long as true effects are not too large. You could then refer the reader to the more general equation by Pustejovsky and point out that it has the deficiency of depending on d, which in turn would be an even larger concern, since it would induce an automatic correlation between the estimates and variances, which you clearly want to avoid.

We thank Reviewer 3 for their prudence in allowing us to leave closed this particular can of worms. We have added an appropriate footnote on page 18, as requested. We prefer to cite the Borenstein (2009) citation used by Pustejovsky rather than citing Pustejovsky directly, as this reviewer’s comment initially caused us a great amount of confusion, as we thought the complaint regarded the need for an assumed value of *w*.

A few other points:

p. 10: "Egger's weighted regression test (Sterne & Egger, 2005) inspects the degree and statistical significance of the relationship between sample size and effect size."

To be precise, Egger's regression test examines the relationship between the effect sizes and their standard errors, not sample size. Yes, the latter is inversely related to the former, but they are not the same.

We’ve fixed this.

p. 11: "If the funnel plot is asymmetrical, the procedure 'trims' off the most extreme study and imputes a hypothetical censored study reflected around the funnel plot's axis of symmetry (e.g., an imputed study with a much smaller or even negative effect size estimate). Studies are trimmed and filled in this manner until the ranks are roughly equal."

The algorithm underlying the trim and fill method can trim off (and fill in) not just the most extreme study, but multiple studies at once. And the "ranks" of what?

p. 12: "In PET, a weighted linear regression is fit to describe the relationship between effect size and precision, as in the Egger regression test. Unlike Egger's test, which considers the slope of this regression, PET considers the intercept of this regression. This represents an extrapolation from the available data to estimate what the effect would be in a hypothetical study with perfect precision."

To be precise, PET (and Egger's regression test) fits a model to describe the relationship between the effect sizes and their standard error, which is a measure of \*imprecision\*, so that the intercept then represents the predicted effect when imprecision is at zero (i.e., when precision is perfect).

Fixed.

p. 13: "PEESE fits a weighted quadratic relationship between effect size and precision."

See above.

Fixed.

p. 19: "PET was performed by fitting a weighted-least-squares regression model predicting effect size as a linear function of the standard error with weights inversely proportional to the square of the standard error. PEESE was also performed, predicting effect size as a quadratic function of the standard error and using similar weights. All meta-regressions were performed using the metafor package for R (Viechtbauer, 2010), using the rma() function to fit a weighted random-effects model with an additive error term."

PET (Egger regression) and PEESE are also in principle meta-regressions, so this could be confusing to some readers. In particular, the models used for PET (Egger regression) and PEESE use a multiplicative term for overdispersion, while the random-effects (meta-regression) model uses an additive term. When the authors refer to "meta-regressions", I assume they are not referring to PET/Egger/PEESE, but only to their moderator analyses. But as written right now, this is confusing.

No, we meant the PET, PEESE, and Egger tests when we said “meta-regressions.” We make this clear now. So these are performed with additive error components, as recommended by [CITATIONS NEEDED].

Also, I looked at moderator\_inspection.R and I see 'rma(yi = Fisher.s.Z, sei = Std.Err, weights = 1/Std.Err, ...'. Why are the authors using 'weights = 1/Std.Err'? This is not how the random-effects meta-regression model is usually fit. This will force the use of inverse standard error weights instead of using the usual inverse sampling variance plus (residual) heterogeneity weights. Also, 'mods = Fisher.s.Z ~ x' is unnecessary. Either use 'rma(Fisher.s.Z ~ x, ...)' or 'rma(Fisher.s.Z, ..., mods = ~ x, ...)'.

p. 25: "In the case of best-practices experiments of aggressive behavior, there was so little residual variance that a confidence interval on I2 could not be calculated. The documentation for metafor suggests that this indicates 'highly (or overly) homogenous (sic) data,' (Veichtbauer, 2010, helpfile for confint.rma.uni) an unusual absence of residual sampling variance. This would be consistent with the presence of bias: effect sizes in this subset seem to reach statistical significance with improbable precision."

Note that a CI \*was\* calculated -- the CI was just equal to the null/empty set (see documentation). Also, I am not sure what the authors mean by "effect sizes in this subset seem to reach statistical significance with improbable precision." Finally, it's "Viechtbauer" (also spelled incorrectly in the References).

Fixed.

p. 28: "CRTT".

I assume the authors are referring to the measurement of "aggressive behavior by allowing participants to administer a painful burst of noise to another participant" (p. 7). But please write out the abbreviation and/or explain this more clearly.

Fixed.

p. 30: "Researchers believe they have well-controlled manipulations yielding robust, unbiased effects. We are concerned that, instead, researchers have poorly controlled manipulations yielding uncertain effects overstated through research bias."

I should have commented on this in the first version. I don't see how any of this follows from the analyses done in the present paper (what researchers believe, whether manipulations are well-controlled or not, whether researchers are overstating their findings).

We have added citations that exemplify how researchers think their manipulations are well-controlled when they are not. We have added a citation for the argument that the research literature is unbiased.

Figure 3: "One not-best-practices cross-sectional study may be an outlier (Sigurdsson, Gudjonsson, Bragason, Kristjansdottir, & Sigfusdottir, 2006, z = 0.49)."

I do not see a point at 0.49 in the top right figure.

Sorry – the z-score is .53, not .49.

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

===============

Reviewer #4: The revisions have greatly improved this commentary on Anderson et al. (2010). Our understanding of issues like publication bias, p-hacking, and outcome switching has increased considerably over the last five years and several of the methods used in the current submission were not around when Anderson et al. published their results. Given the widespread attention and clear relevance of this topic, it is crucial that these rigorous reanalyses are published. The reanalyses are well executed and the results are described well and in a balanced manner. The publication bias present in the database is quite clear from all the analyses presented and so the conclusion that this literature is biased is clearly warranted. I frankly could not think of an alternative explanation for these results other than the ones already discussed by the authors and rightly dismissed. Psychological Bulletin readers need to be made aware of the current results because they cast serious doubt on the conclusions drawn by Anderson et al. (2010). One cannot really blame Anderson et al. for not seeing what we now are much more aware of, namely the "hot" fields that allow flexibility in design and analysis of data and commonly use small samples are at risk of showing effects that are either extremely inflated or might not even be there. Clearly, more research with rigorous methods (large samples and preregistrations) is needed to understand the effects of violent video games on aggression.

[When we read the wise and honorable words of Reviewer 4, we are moved.]