Dear Dr. Johnson,

We are pleased to submit for your consideration a revised form of our manuscript BUL-2015-0509, “Overestimated Effects of Violent Games on Aggressive Outcomes in Anderson et al. (2010).”

The reviews and editorial comments we received were thorough and constructive. The chief concerns might be summarized in three parts: **I**. the calculation of effect size estimates, **II**. attention to heterogeneity, and **III**. considerations of study quality. We address these broad concerns first, with smaller, specific changes discussed later in the replies to individual reviewers.

**I.** There were three separate concerns about the effect-size measures:

A. Were any outcomes dichotomous and being treated as continuous? If so, as pointed out by McGrath and Meyer (2006) base rate differences in outcome can cause differences between *r* and *d*. However, the outcomes here are either continuous or modeled by original authors as roughly continuous (e.g. scales, quantities of hot sauce or noise, counts of aggressive word completions or reaction times to read aggressive words.) Thus, base rate concerns for dichotomous outcomes do not influence our effect sizes measures at all. We now note in our revision that all outcome measures are either continuous or modeled as a continuous. Please see p. xx.

B. There was concern that the use of *r* presented problems in assessment. One issue is that some correlations are point-biserial while others are Pearson moment. Reviewer 3 suggests that these should not be aggregated or compared. We agree, and do not aggregate these effect sizes. We treat experimental studies and correlational studies separately. Therefore, there is no mixing of the Pearson moment correlations from cross-sectional studies and the point-biserial correlations from experimental studies. We note this state in our revision on p xx.

C. Perhaps the most salient argument is that the variance of Fisher’s *z* is not as we estimated, and, therefore, the funnel plots do not accurately represent the relationship between effect size and precision. Such an argument has the potential to threaten the validity of the message of the manuscript, and as such, we put a lot of time into exploring it, and we spend some time writing about it here. Please consider the following:

The design of the experimental work is that a continuous outcome, aggression, is studied as a function of people playing violent and nonviolent video games. Although *d* is clearly preferred for conceptual clarity, we use *r* because the original Anderson et al. paper did so. Use of *r* allows direct comparison of our results to Anderson et al.’s, which is critical. To draw funnel plots, we transformed these point-biserial correlations to Fisher *z*-scores, using the equation 1/sqrt(N-3) to approximate the standard error of *z* (Borenstein et al., 2009, p. 42, equations 6.2, 6.3). This approach is identical to that taken by Anderson et al. (2010). However, if this approximation of the standard error is incorrect, our analyses would be invalid.

We provide two supplementary files to support the efficacy of this approximation. First, we conduct a small simulation. We simulated aggression outcomes from hypothetical violent-video game and nonviolent video game conditions. We then computed Cohen's *d*, transformed it to a point-biserial correlation, and, finally took the Fisher z, all using the standard formulae in Borenstein et al. (2009). We did this simulation 10,000 times for a variety of sample sizes and true effect sizes. For each 10,000 cycle run, we calculated the empirical standard deviation of Fisher's *z* and compared it to the two theoretical approximations: the current 1/sqrt(*N*-3) in Borenstein et al. (2009) and the proposed Equation 16 in Pustejovsky (2014).

As you can see, the two equations differ little in their estimations of the standard error of Z, particularly for sample sizes of about 80 and greater. Both equations seem to overestimate the precision of Z when cell sizes are very lopsided, as they might be in a 3-vs-1 complex contrast.

(As an aside, we note that the Comprehensive Meta-Analysis software is generally pretty carefree about which variance estimator is used for Fisher’s Z. Some macros for entering effect sizes will use the 1/sqrt(N-3) approximation, while others will use the equation suggested by Reviewer 3. It seems that either approximation is generally considered of sufficient quality that it can be used without notice.)

Second, we planned to spot-check a random sample of 15 experimental studies to investigate the magnitude of difference in the standard error across the two variance-estimation methods. This proved difficult. Most articles did not report the per-cell sample sizes. One paper’s reported sample size did not match the degrees of freedom of its test statistic. We had difficulty determining how Anderson et al. extracted a single effect size from other manuscripts. [SO WHAT, JOE?]

**II.** Regarding heterogeneity, we have made thorough additions to the manuscript to better handle and report heterogeneity. Models were changed to use random-effects estimators when possible, the method of random effects estimation is reported, and the I2 statistic and its 95% CI are reported for all models. Heterogeneity does not seem to affect our central argument that violent-game effects on behavior are overestimated. Indeed, studies in the best-practices subset may be *excessively homogeneous* after adjusting for small-study effects, having unusually little residual sampling variance. This is again consistent with a combination of analytic, publication, and/or selection biases: effect sizes in the aggressive behavior best-practices subset seem to reach the .01 < *p* < .05 region with uncanny regularity.

**III.** We have carefully considered the role of study quality in this literature. Despite this consideration, however, we think it best not to make too many changes from the original Anderson et al. codings. Our analyses are intended as a direct reexamination of the analyses and arguments provided by Anderson et al. We are concerned that performing a new analysis based on our own assessment of quality would seem as though we are making *post hoc* changes to the data to attempt to support a pet hypothesis. An earlier draft of the manuscript outlined inconsistencies in the original application of best-practices inclusion criteria, but we felt that such a discussion would lead to an intractable debate about the subjectivities of inclusion criteria. We do exclude two studies due to concerns over relevance (Panee & Ballard, 2002; Graybill et al., 1985).

Moreover, we feel there is little to consider in terms of quality that Anderson did not already measure. Studies are already separated by study design (experimental, nonexperimental). Anderson et al. already consider the validity of the interventions and outcomes in their best-practices criteria. Other indices of quality show very little variance. For example, we are not aware of any studies that attempt to reduce bias by blinding the experimenter to the condition. Still other indices cannot be evaluated due to inconsistent reporting. For example, we cannot assess the influence of participation statistics due to failures of deception, because such failures of deception are not universally reported.

Thus, we feel that Anderson et al. have generally done an adequate job of investigating potential moderating effects of study quality. We disagree with some of their finer points, as we feel that their conceptualization of validity in measurement and manipulation appears to vary from study to study. Moreover, it seems to covary with statistical significance, leading sometimes to the appearance of stronger bias in the best-practices subset than the full set. Again, we prefer not to argue about the inclusion or exclusion of single studies, but rather to evaluate Anderson et al.’s conclusions given their own data.

A thorough reply to individual comments not covered above follows.

Editor:

1. The Editor mentions the McGrath and Meyer paper. We provide a reply above in **IA**.

2. The Editor asks about heterogeneity. We provide a reply above in **II**.

3. The Editor asks us to elaborate on our conclusions regarding publication bias, particularly in light of the appearance of some non-significant findings in the published literature. We consider publication bias to include not only journals’ preference for significant results, but the tendency of researchers to submit for consideration only the significant results. We expect that publication bias is probabilistic, rather than absolute; null results will therefore sometimes be published, albeit less frequently than are significant results. Furthermore, significant findings appear in prestigious journals (*Psychological Science, Journal of Personality and Social Psychology*) whereas null results are published in less prestigious journals (*Journal of Applied Social Psychology*, *Journal of Broadcasting and Electronic Media, Criminal Justice and Behavior,* or Japanese conference proceedings). Researchers may be incentivized to find significant results for a more prestigious publication. Finally, in the case that a student’s post-graduation job does not expect scholarly achievement, we suspect that significant results are nonetheless more likely to be inherited and prepared for publication by advisors, post-docs, or other grad students. Null results are more likely to be file-drawered and forgotten.

4. The Editor asks that we shorten the sections on techniques. We do so (see p. xx), and we also removed two figures (Figures 1 and 2 in the previous ms).

5. The Editor asks that we carefully watch our tone. We have made many small changes to try to make the paper as polite and palatable as possible.

Reviewer 1:

1. Reviewer 1 says that he is sure there are some biases in the literature, but that the biases are likely not strong enough that there is no relationship between violent game exposure and aggression at all. It was not our intention to argue that the *true effects* are necessarily minimal; rather, we meant to say that the *adjusted* *effect size* *estimates* are close to zero. We do not mean to argue that this is strong evidence that there is no effect. Instead, we interpret this as a sign of uncertainty and the need for a strong preregistered replication effort. We thank Reviewer 1 for pointing out the inconsistencies in our language and have made appropriate edits, most saliently to the abstract.

2. Reviewer 1 suggests that the PEESE estimate (*r* = .15) is close enough to the Anderson et al. estimate (*r* = .21) that the results are not remarkable. We disagree for the following reasons. First, the PEESE estimate Reviewer 1 considers is the most optimistic estimator of the effect and is known to be upwardly biased when the null hypothesis is true. Thus, PEESE may overestimate the effect. Second, even if the null hypothesis is false, a decrease from *r* = .21 to *r* = .15 implies a twofold loss of explanatory power (*R*­2 falls from 4.4% to 2.3%) and the need for a twofold increase in sample size (80% one-tailed power at *n* = 136 vs. *n* = 270). Thus, even if violent games do influence aggressive behavior in experiments, the current study offers crucial information for power analysis. We now make note of this in the manuscript (p. xx).

3. Reviewer 1 suggests that our results are not credible given that they are inconsistent with theory. We recognize that the current findings would be inconsistent with contemporary theories of human aggression and media influence. However, to reject the evidence because it does not support the theory would seem to be an unfortunate inversion of the scientific method.

Reviewer 2:

1. Reviewer 2 found we made an unqualified claim that PET, PEESE, and *p*-curve “provide better adjustments” for publication bias. We have softened such claims (see p. xx).

2. Reviewer 2 asked that we pay closer attention to outcome-switching as a form of *p*-hacking. We have elaborated a bit more on the problem of outcome-switching, as requested. (p. xx) Indeed, the flexible quantification of the Competitive Reaction Time Task (see Elson et al., 2014) may constitute a form of outcome switching.

3. Reviewer 2 suggested that we do more to model and describe heterogeneity, as well as point out the influence heterogeneity can have on adjustments for small-study effects. See our reply above in **II**.

4. Reviewer 2 pointed out that our criticism of trim-and-fill as being simply based on effect sizes was inaccurate. We have omitted this unjustified criticism. (p. xx)

5. Reviewer 2 suggested that we apply the Vevea and Hedges (1995) selection model and consider its estimates. The first author lacks the degree of expertise necessary to apply this method. He found a Shiny app online (https://vevealab.shinyapps.io/WeightFunctionModel/), but it crashed when given the dataset. We fear that to learn and implement the Vevea and Hedges selection model would require an excess of time and effort. Therefore, we mention the Vevea and Hedges selection model alongside the Guan and Vandekerkhove Bayesian selection model as potential future directions.

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, we understand the use of power analysis to be quite uncommon in 2010 and earlier. Reforms that placed an emphasis on power calculation did not happen until some years later, e.g. Simmons, Nelson, & Simonsohn’s (2012) 21-word solution. Second, we doubt researchers could have known the causes of heterogeneity in effect size. The Anderson et al. (2010) meta-analysis looked for many suspected moderators of the effect (e.g., sex, age, game perspective) and found none. Additionally, if Anderson knew what moderators would permit acceptable statistical power at both N = 515 (Anderson, Gentile, & Buckley, 2007) and N = 39 (Bartholow & Anderson, 2002), then we expect he would have mentioned such moderators in the 2010 meta-analysis. Simply put, such precision *in a priori* power analysis would seem to require knowledge still unavailable to researchers. Finally, whereas this combination of heterogeneity and power analysis will create an asymmetrical funnel plot, it will still lead to a right-skewed *p*-curve, as *p*-curve is a function of statistical power alone. It is therefore sobering that the *p*-curve for effects on behavior in experiments is essentially flat. We have uploaded a simulation demonstrating this to the OSF repo at <https://osf.io/y2jc6/>. The manuscript now explains why we think heterogeneity and power analysis are not likely to explain the small-study effects (p. xx).

2. Reviewer 3 comments that the cross-sectional literature also shows signs of small-study effects, despite our conclusion that they are “relatively robust and unbiased.” We now describe the pattern of results more thoroughly – we find evidence of small-study effects, but adjusting for these effects does little to change the effect size estimates.

3. Reviewer 3 asks that we not attempt to influence the reader’s priors by alluding to publication bias as being omnipresent. We now omit this rhetorical argument. (p. xx)

4. Reviewer 3 points out that publication bias is threatening for all forms of scientific review, not just meta-analysis; that funnel plots do not summarize the quality of a meta-analysis; that PET meta-regression is not similar to, but rather identical to, the Egger regression; and that funnel plot symmetry requires the absence of both small-study effects *and* heterogeneity. We have amended these statements.

5. Reviewer 3 noticed we cannot support the claim that “dissertations likely represent a minority of all studies conducted on violent games.” We now omit this unsubstantiated claim.

4. Reviewer 3 was concerned about the nature of the effect sizes being synthesized. First, we clarify that we ignored the partial effect sizes as collected by Anderson et al. We were most interested in the effects in experiments, for which the “best partials” estimates are the same as the “best raw” estimates in Anderson et al. (2010). The “best partials” in cross-sectional studies are difficult to interpret and the source of much controversy (see, e.g., Ferguson, 2015; Rothstein & Bushman, 2015). We thought it best not to analyze these. We now make this decision explicit; see p. xx.

5. Reviewer 3 says we were using the wrong method to calculate *t* and p values. We now use the methods suggested by Reviewer 3. All analyses are updated in light of these new t-values.

6. Reviewer 3 asks whether Anderson et al. used Fisher’s Z in their analyses. They did; our manuscript now points that out.

7. Reviewer 3 asks how we computed the SEs for Anderson’s partial correlation coefficients. We did not consider the partial correlations in this study. We now make this clear in the manuscript.

8. Reviewer 3 was concerned that we were combining product-moment and point-biserial correlations. We have clarified that we do not; see our reply above under **II**.

9. Reviewer 3 suggests that we have misestimated the variance of Fisher’s Z. See **IC** above**.**

10. Reviewer 3 asked about the role of Cohen’s *d* in our statement, "P-curve estimates were similarly converted from Cohen's d to Pearson r for consistency of presentation." We now explain that the p-curve code returns estimates in terms of *d*, so we had to convert the estimates to *r*. Our equation for this transformation is given; see p. xx.

11. Reviewer 3 asks about outliers in the funnel plots, particularly Figures 3 and 5. These outliers are now more clearly identified so that the reader can examine their role in the leave-one-out sensitivity analyses.

12. Reviewer 3 asked what we meant by “mixed” statistical significance in a dissertation. We have replaced this confusing term. We meant to say that these studies have several outcomes, only some of which found statistical significance. We now refer to these studies as having all significant outcomes, some significant outcomes, or no significant outcomes. (Table XX)

Reviewer 4:

1. Reviewer 4 points out that others besides Simonsohn et al. have pointed out that trim and fill may perform poorly. We have added further citations.

2. Reviewer 4 requests that we consider other potential moderators and make sure that they are not responsible for the observed small-study effects. We uploaded a supplementary file that inspects the correlations among moderators. This file also makes graphs to visually inspect for possible moderator accounts of funnel-plot asymmetry. Like Anderson et al. concluded, we do not think there are significant moderators of study features that explain differences between effect sizes. Moreover, for experiments of aggressive affect and aggressive behavior, we find little heterogeneity remains after adjusting for small-study effects, implying that there is little residual variance to explain. For more on heterogeneity, see **II**.

3. Reviewer 4 points out that the Egger (1997) regression has been replaced by a weighted regression (Sterne & Egger, 2005). We were not aware of the distinction; we were using a weighted regression all along per Sterne and Egger (2005). We have amended the citation.

4. Reviewer 4 suggested that we change our funnel plots to be centered at zero with shaded contours to emphasize the .01 < p < .05 region. We have adopted this style for our funnel plots.

5. Reviewer 4 requested that we add p-uniform estimates and the Test for Excess Significance. We now provide these analyses and their results.

6. Reviewer 4 asks that we post the raw data. At the time of submission, we had asked Dr. Anderson for permission to share the data, but had not yet received a reply. He has since graciously agreed to sharing the data, which has now been added to our online repositories.