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. The claim is that these should not be aggregated or compared. We agree---they should not. In fact, we do not. We treat separately experimental studies and correlational studies. We have now gone back and confirmed that in all experimental studies the effect sizes are indeed point-biserial while in all correlational studies, the effect sizes are indeed Pearson moment. Therefore, there is no mixing of the two different types. We note this state in our revision on p xx.

C. Perhaps the most salient argument is that the variance of *z* is not what we think, 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. For the following reasons, we now believe the argument is not right, and that our treatment of variance, though not exact, is quite accurate. 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 and used the Borenstein approximation of the variance. If this approximation is either wrong, too coarse, or inapplicable, our work would indeed be threatened.

The approximation is excellent and applicable. Perhaps the quickest and most persuasive way of showing is through 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. 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 standard deviation of Fisher's *z* and compared it to the theoretical value, 1/sqrt(*N*-3), that we use to construct funnel plots. As you can see, the correspondence between the theoretical values and simulated values is quite good. There is a slight discrepancy for small sample sizes with large effects, but overall the Borenstein theoretical approximation is more than adequate.

Reviewer 3 cites the argument of Pustejovsky (2014) as rationale for concern. Pustejovsky’s is an excellent paper, but in our opinion, his treatment of controlled experiments on p. 96 is non-standard and objectionable. He is trying to treat *r* as coming from a slope measure with units in both the numerator and denominator. As such, he requires the investigator to put units and measure on the independent variable, in this case, whether the participant was in a violent or nonviolent video game condition. Of course, this is silly; there is neither a unit-of-measure nor any value; one cannot assign a value to Pustejovsky's omega in his Equation (4). We think the appropriate approach is to dispense with the interpretation of *r* as corresponding to a slope and treat it as related as an algebraic transform of a measure of accounted variance. For this interpretation, the Borenstein approach does exceedingly well as we have shown in our small simulation. Note that we do not mix interpretations because we treat experimental (biserial, categorical independent variable) and correlational data (Pearson moment, continuous independent variable) separately.

As an aside, the senior author, Jeff Rouder, is sensitive to the issue. He develops Bayes factor solutions for common designs. Rouder's group recommends scaling the priors on slopes by the covariance of the dependent measures in regression contexts but not in ANOVA contexts. This different treatment reflects that the fact that continuous covariates have a unit of measure while categorical ones do not. (see Rouder et al., 2012, J. Math Psych; Rouder and Morey, 2012, Mut Var Beh Res).

The concern about the chain of transforms, their meaning, and the variance of Fisher's Z has led to the following revisions. [REVISIONS GO HERE]

**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.

**III.** Regarding the quality of studies, 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 look 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. Futhermore, there is little to consider in terms of quality that Anderson did not already measure. For example, we are not aware of any studies that blind the experimenter to the condition. Thus, we prefer to re-analyze the arguments of Anderson et al. on the basis of their own data. [JOE, What did we do?] [GOTTA MENTION SOME KIND OF REVISION]

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 expand our conclusions regarding publication bias. 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).

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 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). This is practical knowledge, even if the null is false. 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 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)

4. Reviewer 2 suggested that we do more to model and describe heterogeneity. See our reply above in **II**.

5. Reviewer 2 suggested that we apply the Vevea and Hedges (1995) selection model and consider its estimates. The first author admits that he lacks the degree of expertise necessary to apply this method. He knows nothing of the Newton-Raphson method. He found a Shiny app online, but it crashed when given the dataset. Given the challenges involved in applying the Vevea and Hedges model, we leave it as an exercise for the reader. We mention the Vevea and Hedges method alongside the Guan and Vandekerkhove method as potential future directions.

Reviewer 3:

1. Reviewer 3 asks that we not attempt to influence the reader’s priors by referring to the omnipresence of publication bias. We now omit this argument. (p. xx)

2. 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.

3. 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 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**.

6. Reviewer 3 suggests that we have misestimated the variance of Fisher’s Z, and that we should be using the equations given in Pustejovsky (2014). We have given very careful consideration to the Pustejovsky (2014) manuscript but found it irrelevant to the current manuscript. We note that *w*2 does not represent as suggested, but rather the assumed change in the predictor *x* inflicted by the experimental assignment. For more, see **IC** above**.**

~~In correlational research, researchers inspect the correlation between some aggressive outcome (e.g., behavior) and some measure of chronic violent game exposure (e.g., hours of violent video games played in the past 5-6 years, Anderson & Dill, 2000). In experiments, researchers inspect the relationship between an aggressive outcome and a brief measure of violent game exposure (e.g., fifteen minutes with a violent game). The two paradigms consider very different mechanisms and make very different assumptions about the underlying relationship.~~

~~But if one were to use Pustejovsky’s equation 4 to synthesize these results, we submit that the results would be insensible. Recall that Pustejovsky’s equation suggests the assumption that observations in the control group have~~ *~~x~~~~i~~*~~=~~ *~~q~~* ~~for~~ *~~i~~* ~~= 1, …,~~ *~~n~~~~1~~*~~, whereas the observations in the treatment group all have~~ *~~x~~~~i~~* ~~=~~ *~~q + w~~* ~~for~~ *~~i~~* ~~= 1, …,~~ *~~n~~~~2~~*~~. Here~~ *~~w~~* ~~indicates the displacement in~~ *~~x~~* ~~from the control mean~~ *~~q~~*~~. If we were to consider~~ *~~x~~* ~~some measure of months or even years of violent game exposure, as used in correlational research, the displacement~~ *~~w~~* ~~inflicted by 15 minutes of play must be perishingly small. (Note that~~ *~~w~~*~~2~~ ~~does not represent as suggested, but rather the assumed change in the predictor~~ *~~x~~* ~~inflicted by the experimental assignment.) As~~ *~~w~~* ~~tends towards zero, the equation~~ *~~r~~~~ce~~* ~~= tends towards 1.~~

~~Under these assumptions, if fifteen minutes of violent game exposure can inflict any change in aggressive behavior, then the effect of fifty hours or more over some years would be nothing short of titanic. These would seem to reflect a fundamental “apples and oranges” problem that forbids direct comparison of the experimental and cross-sectional phenomena synthesized in this literature. We think that this is a valuable and interesting insight, but feel that it is beyond the scope of the current manuscript. We respectfully request that we do not use Pustejovsky’s equations for estimating z~~~~ce~~ ~~or its standard error.~~

7. 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.

8. 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).

9. 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.

10. 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 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.

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

3. 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.

4. Reviewer 4 requests better attention to heterogeneity. As requested, we report I2 and its CI; see **II**.

5. 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.

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.