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 are distortions. 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 exceedingly 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 Anderson et al., our target paper, did so. Consequently, we can compare 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 Bornstein 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 very small sample sizes, but overall the Borenstein theoretical approximation is more than adequate.

Reviewer 3 cites the argument of Pustejovsky (2014) as rationale for concern. Pustejovsky 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.

**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?

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:

We ask Reviewer 1 to reconsider. The PEESE estimate Reviewer 1 considers is the single most optimistic estimator of the effect and is known to be upwardly biased when the null is true. Even so, Reviewer 1 may find it important that 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).

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.

We recognize that the current findings would be inconsistent with contemporary theories of human aggression. 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:

Reviewer 2 made a number of valuable suggestions where our rhetoric grew sloppy, e.g., in our unqualified claim that PET, PEESE, and *p*-curve “provide better adjustments” for publication bias. We have softened such claims and used more precise rhetoric as advised.

We have elaborated a bit more on the problem of outcome-switching, as requested. Indeed, the flexible quantification of the Competitive Reaction Time Task (see Elson et al., 2014) may constitute a form of outcome switching.

We have added a section cautioning the reader as to the potential influence of heterogeneity in meta-analysis. We caution the reader that such heterogeneity can also influence conclusions drawn from meta-regression and p-curve techniques. We omit the unjustified criticism of the trim-and-fill technique.

[TODO: REVIEWER 2 ASKED THAT WE TALK ABOUT AND APPLY THE VEVEA & HEDGES 1995 METHOD. I DON’T REALLY WANT TO DO THAT – THE PAPER IS ALREADY HUGE. I ALSO CAN’T UNDERSTAND THE DAMN METHOD. I FOUND A SHINY APP ONLINE AND IT DOESN’T WORK WITH OUR DATA SET.]

We have researched the Vevea and Hedges (1995) method but are having considerable trouble performing it. I know nothing of the Newton-Raphson method and cannot implement it. We found an online Shiny application for performing the Vevea and Hedges method, but it only yielded error messages when given our data. Given the considerable challenge in implementing this method, we respectfully suggest that it is beyond the scope of the current manuscript.

As requested, we now use random-effects models wherever possible. We report our choice of random-effects estimator (restricted maximum likelihood) and the I2 statistic and its confidence interval.

Reviewer 3:

Reviewer 3 similarly indicated a number of places where our rhetoric was too strident. We thank Reviewer 3 for these constructive critiques and have amended our language as appropriate.

We have clarified 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 touch these.

Similarly, we do not combine Pearson product-moment correlations and point-biserial correlations. We now make this clearer in the revised manuscript.

Also with attention to the computation and synthesis of effect size estimates, we thank Reviewer 3 for bringing the work of Pustejovsky (2014) to our attention. Because the original Anderson et al. meta-analysis was performed with Borenstein’s software, Comprehensive Meta-Analysis, we had attempted to mirror its approach as closely as possible. This lead us to adopt the Borenstein perspective that, as Pustejovsky says, frames conversion formulas as “applying across entire categories of effect size measure and as being reversible […] as if it were an algebraic identity.”

There are a number of reasons we are reluctant to implement the refinements suggested by Reviewer 3. First, it is our impression that the concerns raised by Pustejovsky apply chiefly when attempting to synthesize correlational designs with experimental designs. As we maintain separate meta-analyses for controlled experiments and cross-sectional designs, Pustejovsky’s suggestions would appear to be relevant only when trying to interpret the relative magnitude of the synthetized effects in experiments against those in cross-sections. We avoid making such interpretations in our manuscript.

Second, the controlled-experiment correlation estimator proposed by Pustejovsky does not seem appropriate to the current research. 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 *xi* = *q* for *i* = 1, …, *n1*, whereas the observations in the treatment group all have *xi* = *q + w* for *i* = 1, …, *n2*. 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 *rce* = 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 zce or its standard error.

We changed our method of t-value and p-value calculation as requested by Reviewer 3. All analyses are updated in light of these new t-values.

We wish to point out some reasons why we think it unlikely that the funnel plot asymmetry is caused by a combination of heterogeneity and *a priori* power analysis. 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 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. 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/>.

Outliers are more clearly identified in Figures 3 and 5 so that the reader can examine their role in the leave-one-out sensitivity analyses.

We have replaced the confusing term “mixed” statistical significance. We meant to say that these studies have several outcomes, only some of which found statistical significance.

Reviewer 4:

We were previously using the weighted regression, unaware of the distinction between the Sterne & Egger (2005) method and the Egger (1997) method. Thus, the results have not changed, but the citation has.

Per the reviewer’s suggestions, we have added *p*-uniform and the Test for Excess Significance. We have also switched to zero-centered, contour-enhanced funnel plots to better illustrate the p = .05 threshold.

As requested, we report I2 and its CI. A supplementary file inspects the correlations among moderators and visually inspects 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.

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