Dear Dr. Johnson,

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

In your letter, you raise a number of big-picture concerns, which we summarize as three questions: **I.** How should these results be interpreted? **II.** Do our decisions to exclude studies cause their own bias? and **III.** What are the implications for theory? We discuss these points below.

**I. “How to best interpret your Results?”**

A. In your letter, you ask us to clarify how best to interpret the results. We have clarified accordingly. Our results can be summarized in brief as follows:

There is clearly a large degree of bias among the collection of experiments of aggressive affect and aggressive behavior. In contrast, bias seems absent from experiments of aggressive cognition and of physiological arousal. It also seems absent from cross-sectional, correlational studies.

We think the large degree of bias in the above two subsets has substantial, important theoretical and practical ramifications. We discuss here the practical ramifications; theoretical ramifications are described below in response to **III.**

From a practical perspective, the question of whether there are violent-media effects seemed to be solved prior to our work. Most researchers in the field are strongly convinced of these effects, so much so that, to these experts, skepticism sometimes seems to be an act of willful deceit. For example, Strasburger, Donnerstein, and Bushman (2014, p. 572) write, “Despite thousands of research studies on media effects, many people simply refuse to believe them. […] Of course, many people still believe that President Obama wasn’t born in the United States, President Kennedy wasn’t assassinated, men didn’t walk on the moon, and the Holocaust didn’t occur.” Researchers also believe that the evidence is very strong, very compelling, and very robust. A research summary published last year by Anderson, Bushman, Donnerstein, Hummer, and Warburton (2015, p. 7) claims “Media violence research has provided one of the largest and most well-understood bodies of scientific evidence in all of social and behavioral science.” Anderson and Bushman (2002, p. 2377) report that violent-game effects are “larger than the effects of calcium intake on bone mass or of lead exposure on IQ in children,” and the American Academy of Pediatrics says in their 2009 policy statement on media violence that “the strength of the association between media violence and aggressive behavior […] is nearly as strong as the association between cigarette smoking and lung cancer – associations that clinicians accept and on which preventative medicine is based without question. (p. 1497)”

The best piece of evidence for this strong stance is the Anderson et al. meta-analysis. Our results are critically important because they bring this strong claim into question in two distinct ways. First, we conclude that there is not as much evidence for effects as claimed, primarily because in the face of such bias it is hard to trust the data. Second, we conclude that effect sizes are not as large as claimed. Scientists and professional associations need to be aware that the evidence for these phenomena is tenuous rather than firm. Like Reviewer 4, we feel that it is crucial that the scientific community be aware of these limitations of the literature and adjust their research practices, rhetoric, and policy recommendations accordingly.

As to whether there is or is not “an effect” in experiments, we think in the face of such large bias it is premature to try to come to a conclusion. We report adjusted effect size estimates more as a courtesy, but note that in the face of such a problematic bias, definitive statements are difficult. For us, the most dramatic findings are the funnel plots themselves, which are alarming and directly contest the strength of evidence for purported violent-game effects in experiments. This is particularly true of the Anderson et al. (2010) best-practices subset, which seems to have preferred the inclusion of statistically-significant results.

REVISIONS: We now explain in the discussion (p. 29) and summary (p. 39) that the present results indicate researchers should be less confident in violent-media effects than before, and suggest that policy statements consider taking a less strident tone. Additionally, we have changed the title to reflect the overstatement of evidence, rather than the overestimation of effect sizes, in Anderson et al., (2010). Additionally, we have edited the discussion to be more interpretation, less statistics.

B. In interpreting the results, you ask us to provide greater nuance than was available in the previous abstract and title, given that not all analyses indicate overestimation of effects.

REVISIONS: We have changed the title, abstract, and discussion (p. 29) to reflect that the bias was specific to experiments studying aggressive affect and aggressive behavior.

C. By way of nuance, you mention that the cross-sectional effects are unbiased. You also mention the exclusion of the cross-sectional “best partial” results from discussion as a potential weakness.

Cross-sectional studies do not seem to suffer from bias and the results from this class of experiments may be on firmer ground than the experimental studies. One limitation of these studies is that they are correlational and do not provide evidence of a causal relationship per se. Partial correlations can address this to a degree. In other research, partial correlations in cross-sectional and longitudinal research are reported as being quite small (Anderson et al., 2010; Ferguson, 2015; Furuya-Kanamori & Doi, 2016). In general, we find it plausible that long-term violent game use would have some effects. This is not in conflict with our core finding that the effects of *short*-term violent game use are biased, tenuous, and overstated.

REVISIONS: We emphasize in the abstract the absence of bias from the cross-sectional literature. On page 29, we describe the clear associations indicated by cross-sectional studies, but caution that other research suggests these are small in longitudinal research or when partialing out the effects of potential confounds such as sex. We also make reference to the broader literature (and controversy) regarding partial correlations in cross-sectional media violence studies (p. 30).

D. You ask us further where there is or is not an effect after adjustment for small-study effects. That is very hard to say. The bias is very strong, and so the studies have little probative value. Results, particularly statistical significance, depend substantially on the meta-analytic technique used. In our view, the combination of strong bias, weak data, and diverse statistical assumptions prevents valorization of any single estimator as the one estimator which determines statistical significance or nonsignificance.

For example, you point out that some analyses indicate statistically significant effects, e.g., that the PEESE estimate of effects on aggressive behavior is statistically significant and has no residual heterogeneity. But one must consider also that PET and p-uniform do not find a significant effect. Taken together, the estimators suggest that the evidence is very shaky.

Note that other scholars have suggested less nuanced approaches that may inappropriately favor the null hypothesis. For example, Stanley and Doucouliagos (2013) suggest interpreting a nonsignificant PET estimate as a null result in the PET-PEESE framework. However, we understand that PET has poor sensitivity to effects: it assumes, and favors, the null hypothesis. In similar fashion, we expect that PEESE has poor specificity: it assumes, and therefore may favor, the alternative hypothesis. In general, the previous literature says little about the Type I and Type II error rates of these estimators.

You say that these results and claims are ambiguous, but we prefer to see them as nuanced. Analysis is always a function of models and assumptions, and here we show that the effects are sufficiently biased that assumptions in correction matters. The contamination by bias is so pernicious that the accumulated literature has little probative value, leading the effect size estimates to be highly dependent on the modeling assumptions.

REVISIONS: The results section (p. 26) now mentions the statistical (non)significance of the estimators, but the reader is cautioned that little is known about the error rates of the estimators. Pages 26 and 29 mention the dependence on modeling assumptions.

E. Furthermore, you are curious about potential patterns of moderation and heterogeneity, and how they might influence where and when “the effect” exists.

We are concerned that there is little hope of finding significant moderation: Anderson et al. (2010) found none of their hypothesized moderators predicted the size of the effect, and we are concerned that publication bias may conceal patterns of moderation. Indeed, after adjustment for small-study effects, there is often little residual heterogeneity, and in some cases, a suspicious excess of homogeneity. This also addresses your concerns about the implications of the Van Bavel et al. paper for our manuscript: The problem is not that effects on behavior and affect are sensitive to the context, but rather that these effects are sensitive to nothing but the sample size. After adjustment for small-study effects, there is hardly even sampling variance in the effect size, much less heterogeneity due to context.

REVISIONS: We now very briefly summarize Anderson et al.’s fruitless search for moderators (p. 17) and explain how publication bias may stymie the search for moderation in meta-regression (p. 32).

**II. Do your exclusions cause their own bias?**

A.We excluded a few studies because we found they did not fit reasonable inclusion criteria. You ask if these additional exclusions affect our results. They do not.

First, we have decided not to exclude Matsuzaki et al. (2004). We spoke with another Anderson co-author, Aikiko Shibuya, and she said that there was no mistake in the entered effect sizes, which do not appear in the paper but were instead provided directly from the authors in personal correspondence. Thus, effect sizes from Matsuzaki et al. (2004) now appear in our main analyses. Its inclusion does not change our conclusions, as the cross-sectional literature was already considered robust and unbiased. We note in the supplement that the Matsuzaki et al. effect sizes are unusually large for their sample sizes and have some influence on meta-regression analyses.

To consider how our results would differ from Anderson et al.’s, you must first consider that their trim-and-fill analyses were restricted to the “best experiments” and “best partials” datasets. Only one of the excluded effect sizes, aggressive affect from Panee & Ballard (2002), was included in these analyses.

To show that these exclusions are not influential, we performed naïve fixed-effects and trim-and-fill results, as Anderson et al. did, both with and without our exclusions. The exclusion or inclusion of Panee & Ballard had little influence on our results. With Panee & Ballard included, we get *r* = .294, with a trim-and-fill adjustment of six imputed studies, *r*+ = .242. With it excluded, we get *r* = .289, with a trim-and-fill adjustment of six imputed studies, *r*+ = .238.

Note that we couldn’t quite reproduce Anderson et al.’s trim-and-fill results. Anderson et al. report trim-and-fill results for best experiments *r+* = .294, with zero imputed studies. Without exclusions, we obtain naïve FE result *r* = .289, and trim-and-fill imputes 6 studies to the left, reducing the effect size to *r+* = .238. We are not sure what’s responsible for these differences, but considering how nakedly asymmetrical the funnel plot of best-practices aggressive affect experiment effect sizes is, it seems implausible that trim-and-fill would have imputed zero studies as reported by Anderson et al. (2010).

Not only did the exclusion of Panee & Ballard (2002) have little influence on the results, but it is also clearly irrelevant. The experiment does not manipulate the presence of violence in the video game; all participants played the same violent game. Furthermore, its exclusion reduces, rather than exaggerates, the support for some of our conclusions; thus, we are not being biased by excluding it. It is a high-effect-size, low-sample-size study, and so its inclusion would further indicate funnel-plot asymmetry. Its inclusion caused the PET estimate of all-experiments aggressive behavior to fall slightly (r = .10) and caused best-practices aggressive affect to fall further (PET, r = -.14; PEESE, r = .14). However, it has very small p-values, and so its inclusion causes a slight increase in the p-curve estimates for all-experiments aggressive behavior (r = .09) and best-experiments aggressive affect (r = .26).

The exclusion of Graybill, Kirsch, & Esselman (1985) doesn’t influence the results because Anderson et al. coded it as not-best-practices, thereby excluding it from their trim-and-fill analyses. As you suggested, we looked for effect sizes that were not manipulation checks. Anderson et al. entered two: *r* = -.02 and *r* = .02, N = 116. They are an unusual measurement of the “direction” and “type” of a child’s aggressive thoughts, rather than the *quantity* of aggressive thought. Thus, they are not relevant to the current research question. However, keep in mind that these are nonsignificant results, and so would not influence *p*-curve or *p*-uniform, and that it is a high-precision, low-effect-size study, and so would tip the scales towards funnel plot asymmetry. Although it would support our argument, we think it best to exclude this study, as it is irrelevant.

REVISIONS: We now include Matsuzaki et al. (2004) and explain that Graybill et al. (1985) had no relevant effect sizes for meta-analysis. The above are explained on pages 21-22. Our attempt to reproduce Anderson et al.’s naïve and trim-and-fill analyses appears in a footnote.

**III. Where is the theory?**

A. You comment that the manuscript does little to provide new knowledge, and that the results are in conflict with all extant theories of aggressive behavior, which predict that “stimuli aggressive beget behaviors aggressive.”

The violent video game literature not only has important policy implications, but it also has important theoretical implications for psychologists’ understanding of aggression. The main theoretical apparatus to explain violent game effects comes from the General Aggression Model (GAM), which posits that exposure to violent video games causes increases in arousal, aggressive affect, and broadly-defined aggressive cognition. These aggressive internal states, in turn, are expected to cause aggressive behavior. Violent media studies are thought to provide a paradigm for testing and elaboration on the GAM, which can then be extrapolated to the effects of real-world violence on aggressive and even violent behavior (DeWall, Anderson, and Bushman, 2011).

How do our results affect the plausibility of the GAM? In our previous draft, we decided not to make any statements. We agree with your comments that we need to do so, even if we make statements that are limited in scope:

Violent video game effects are anticipated by GAM and similar theories. A failure to find them in a large data set should count as contradictory or negative evidence for the theories. In our case, we show that the support for effects is overstated, and, as a consequence, we think the stated support for these theories has been overstated as well.

REVISIONS: We now make this point explicitly in the discussion, p. 29. We describe the mechanisms of violent-game effects as proposed by the GAM and consider how the present results, in combination with other recent studies of similar mechanisms, suggest a wave of rising skepticism. We have added a section to the discussion in which we describe the above implications for theories of aggressive behavior (p. 35-36)

B. An additional theoretical consideration bears mention. A central mechanism of the GAM involves the activation of aggressive thoughts through environmental stimuli; that activation of thought is presumed to lead to increases in aggressive behavior. That is, violent games are thought to change behavior, at least in part, through “priming” (see Anderson et al., 2010, p. 155: “[…] the existing short-term effects are mainly the result of priming effects”). The last six years have seen a number of high-powered failures to replicate other priming effects on social behavior. Thus, these results fit well with the broader literature urging skepticism of social priming phenomena and theories.

REVISIONS: We summarize this broader literature and make this argument on pages 30–32.

C. Theories of aggression (i.e., GAM) tend to assume that violent media cause aggressive affect. We are suspicious of this assumption given the very strong bias among experiments of aggressive affect. In rereading this literature, we noticed that there is generally little theoretical justification for this phenomenon; a mechanism is rarely posited and never tested.

REVISIONS: We suggest that theories of aggression would benefit from greater attention and testing of whether and how aggressive affect is stimulated (p. 34-35).

**Other editor’s concerns**

1. A number of notes on the manuscript indicate a desire for clearer description of the nature, prevalence, influence, and meaning of heterogeneity.

REVISION: We have collected our mentions of heterogeneity into special sections in the introduction (p. 15), results (p. 26).

2. The rationale behind the quadratic term in PEESE needed better explanation.

REVISION: We have elaborated on the rationale behind this quadratic term (p. 12-13).

3. You ask for citations regarding the problems of p-curve and p-uniform under heterogeneity.

REVISION: We cite recent criticism from van Aert et al. (in press); see page 17.

4. You ask why we include the Test for Excess Significance despite thinking it a poor test. We had included it previously due to a reviewer request.

REVISION: We moved this test to the supplement.

5. You suggest that the opening paragraph of the Results could do more to orient the reader in a *Psychological Bulletin* fashion.

REVISION: We have changed the paragraph to better introduce the relevant categories and the number of included effect sizes (p. 22-23)

6. You suggest citation of Kerr’s HARKing paper when describing the problems in analysis of effects on aggressive affect.

REVISION: We have added this citation (p. 34).

**Reviewers’ concerns**

We thank Reviewers 2 and 4 for their kind words. Below, we address the remaining concerns of Reviewers 1 and 3.

Reviewer 1:

1. Reviewer 1 suspects there is likely publication bias in most (social) psychological literatures. In that regard, Reviewer 1 seems to imply the current results are not particularly remarkable. We agree that most literatures in social psychology are probably influenced by publication bias to some degree. It is for this reason that analyses like ours are crucial in diagnosing publication bias for the attention of future research. The literature should not endeavor to be “only as bad as the rest of social psychology”; it should instead endeavor to be *correct*. With this in mind, we must correct Anderson et al.’s conclusions regarding the absence of bias.

2. Reviewer 1 suggests that our conclusions are not credible given the effects of very subtle manipulations in stimulating aggression (e.g., black-color primes, heat-word primes, pictures of guns). If these subtle stimuli can provoke aggression, then the stronger stimuli involved in violent video game play must also provoke aggression.

We have not been able to find the relevant citations for Reviewer 1’s claims. We have found the claim that that teams in black or red jerseys are charged with more fouls (Frank & Gilovich, 1988), but follow-up studies contest this claim (Caldwell & Burger, 2010). Heat-word primes have been linked not to aggressive behavior, but rather to hostile attribution bias (DeWall & Bushman, 2009). However, evidence suggests that this finding does not replicate (McCarthy, 2014). Actual temperature has been linked to aggressive behavior (Anderson, Anderson, Dorr, DeNeve, & Flanagan, 2000), but experiencing hot temperatures is very different from priming with heat-related words. Meier, Robinson, & Wilkowski (2006) report that aggressive lexical primes have no main effect on aggressive behavior, but that they interact with trait agreeableness. However, the evidence for even this interaction is slim: *p* = .044 under one CRTT quantification, and *p* = .262 under another. In general, there seems to be little evidence that subtle primes provoke aggressive behavior.

It is our concern that the problems of p-hacking, HARKing, and selective report may be found in these other literatures on aggression (e.g., the “weapons effect” as described by Berkowitz). Of course, a detailed inspection of those literatures is beyond the scope of the current paper.

REVISIONS: We cite the broader theoretical literature that predicts that violent stimuli should prime aggressive behavior, as well as the recent empirical literature that finds that subtle priming manipulations do not influence behavior (p. 30-32).

3. Reviewer 1 suggests that our results would benefit from more thorough attention to theory. See our response to **I** above.

Reviewer 3:

Reviewer 3’s comments lead to a number of further revisions.

1. We apologize that our simulation was not available as we thought it was, and appreciate Reviewer 3’s having performed a simulation of their own. We appreciate Reviewer 3’s prudence in allowing us to leave closed this particular can of worms.

REVISIONS: 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. (The original comment and citation of Pustejovsky initially caused us a great amount of confusion, as we thought the complaint regarded the need for an assumed value of *w,* per Pustejovsky’s (2014) equation 4.)

2. Reviewer 3 mentions ambiguity around the term “meta-regression” as to whether it means tests for moderators in general or tests for small-study effects in specific (i.e., Egger, PET, PEESE). Furthermore, Reviewer 3 suggests that PET and PEESE use multiplicative error terms rather than additive random-effects error terms.

By “meta-regressions” we did indeed mean the Egger test, PET, and PEESE.

The use of weighted random-effects models with additive error terms, fit by use of restricted maximum likelihood, is consistent with recommendations from Thompson and Sharp (1999). Nevertheless, we have also conducted “dispersion” models (as suggested by Moreno et al., 2009) using weighted fixed-effects models with multiplicative error terms. These were fit using the lm() function in base R.

These results are available in the supplement. In general, they differ slightly from the additive-error results, usually by little more than +/- .02 units of Pearson *r*. Notable differences between the random-effects and dispersion models are that (1) The dispersion models estimate stronger evidence of effects on aggressive behavior in best-practices experiments, and (2) the dispersion models estimate weaker evidence of effects on aggressive cognition in experiments.

Regarding (1), in best-practices experiments of aggressive affect, the PET estimate for the dispersion model is r = .096, an increase of .025 from the random-effects model. Whereas the random-effects model reports a nonsignificant PET estimate, p = .188, the dispersion model reports a significant PET intercept, p = .028. The PEESE estimate in the dispersion model is r = .17, slightly larger than that of the random-effects model, r = .15.

Regarding (2), in best-practices experiments of aggressive cognition, the PET estimate was significant in the random-effects model, as originally reported. In contrast, the dispersion PET model estimates a smaller, nonsignificant effect. Additionally, the random-effects Egger test did not detect significant asymmetry in either the best-practices or full sets of studies, whereas the dispersion model does.

REVISIONS: We explicitly mention the Egger test, PET, and PEESE instead of referring to them collectively as “meta-regressions.” (p. 20). We include the dispersion models in a supplementary table and discuss these results in the supplement, as indicated by a footnote on page 20.

3. Reviewer 3 asks for moderator analyses that do not adjust for small-sample effects. These moderator analyses would be redundant with those performed by Anderson et al. (2010).

REVISIONS: We remind the reader in the text that Anderson et al. (2010) did not find any significant moderators of the effects. We provide the correlation tables of moderators for each group of experiments in the supplement, and describe in the text how publication bias might frustrate the search for moderators (p. 32). We have omitted the footnote, as it is now redundant with the Anderson et al. (2010) search for moderators.

4. Reviewer 3 asks us to support our statement that “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.”

REVISIONS: Although we can support this with citations, we feel this statement is out of place and have decided to omit it.

5. Reviewer 3 points out that the caption to Figure 3 misidentifies an outlier. We’d confused Pearson *r* with Fisher’s *z*.

REVISIONS: The correct effect size is now listed as *z* = 0.53.

6. Reviewer 3 raises a number of other issues regarding language and clarity.

REVISIONS: We now refer to the Competitive Reaction Time Task by name and not by acronym. We refer to the relationship between effect sizes and standard errors in regression, rather than the relationship between effect size and sample size. We attempt to explain the trim-and-fill technique in greater detail, but can only provide what is given in the original Duval & Tweedie (2000, p. 457).