Main points:

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

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

RESPONSE: There is clearly a large degree of bias among the collection of experiments of aggressive affect and aggressive behavior. We think the large degree of bias has substantial, important theoretical and practical ramifications. We discuss first the practical ramifications, then the theoretical ramifications.

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 strongly believe in it, so much so that 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.” 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 need to be aware that the evidence for these phenomena is tenuous rather than firm. 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. REVISIONS: ???

For us, the most dramatic findings are the funnel plots themselves. These 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. 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. REVISIONS: ???

Experiments of aggressive cognition may be less biased and may enjoy a stronger evidence base. However, there is considerable heterogeneity among studies. Since there is little evidence that this heterogeneity is caused by differences in stimuli or study population, perhaps they are due to differences in measurement methodology.

REVISIONS: On page 34, we explain how publication bias might make it difficult to detect moderation.

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.

As for whether a significant effect remains after adjustment for small-study effects, that is very hard to say. Regarding aggressive behavior, PET and p-uniform do not find a significant effect, whereas PEESE and p-curve do. Although Stanley and Doucouliagos (2013) suggest interpreting this as a null result in the PET-PEESE framework, 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. However, the previous literature says little about the Type I and Type II error rates of these estimators. We prefer to interpret the effect size estimates. We would not characterize these results as ambiguous but rather 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: We now make the above point in the discussion, page 29.

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

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. To show that the exclusions have a marginal effect, we created a supplementary table conducting the analyses performed by Anderson et al., which are naïve fixed-effects meta-analyses and trim-and-fill analyses. Their trim-and-fill analyses were restricted to the “best experiments” and “best partials” datasets, in which only one the removed effect sizes was included.

Anderson et al. report trim-and-fill results for best experiments *r+* = .294, with zero imputed studies. 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 we have very high confidence in our procedures and results. Moreover, as a check, we can compare our results to the funnel plots. 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).

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 that the Matsuzaki et al. effect sizes are unusually large for their sample sizes and have considerable influence on meta-regression analyses.

The exclusion of Panee & Ballard (2002) had little influence on the results. In general, we feel that the study is irrelevant, as it does not manipulate the presence of violence in the video game, instead changing the instructions given to participants in an earlier training level. Regarding the effects of this exclusion on the results, 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 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 explain the above on pages 21-22 and provide naïve and trim-and-fill analyses in the supplement.

**III. Where is the theory?**

The violent video game literature has important policy implications regarding the regulation of violent video games. It too has important theoretical implications. The main theoretical apparatus to explain violent game effects comes from the General Aggressiveness 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.

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.

The biggest theoretical issue is about the effect of the violent video games on internal states. There are no theories about what makes properties, features, or attributes of a video game elicit aggressive affective states. Without guidance, there exists the possibility of the following element of unfalsifiability: If one finds violent video game effects, then one might well conclude that the employed video games elicited aggressive affective states. If one does not, one can always fault the game. It wasn’t sufficiently violently efficacious, whatever that means. With consideration of this unfortunate Catch 22, we worry that the the Gam/New Look theories haven’t been specified sufficiently to be falsifiable. REVISIONS: We have written a new discussion about future advances to gain falsifiability.

Put simply, GAM theorizes two causal processes: an effect of the external environment on internal states, and an effect of internal states on aggressive behavior. The latter process is not controversial, as it seems trivial to argue that those who feel aggressively and think aggressively are more likely to aggress than those that do not. The former process, however, is in need of better empirical support. It is not clear that violent games cause aggressive affect, although they may cause certain forms of aggressive cognition. It is also not clear that violent games cause aggressive behavior, suggesting a weak relationship between aggressive cognition and aggressive behavior.

After considerable thought and discussion, we remain uncertain that violent video games create an aggressive affective state, as has been claimed. We suspect that playing violent video games is intrinsically rewarding (otherwise few would choose to do it), and as such may lead to pleasant, rather than hostile or frustrated, affective states. In support of this view, we note that the funnel plot of experimental effects on aggressive affect was the most clearly asymmetrical indicating the most bias in establishing the effect.

REVISIONS: We have added a section to the discussion in which we describe the above implications for theories of aggressive behavior (p. 35-36)

Finally, you ask us to consider the degree to which these effects may be contextually sensitive (e.g., the recent Van Bavel paper in PNAS). The problem posed by our main findings are not that the effects are contextually sensitive; rather, the problem is that they don’t seem to be sensitive to anything but the sample size of the experiment. The evidence base is so badly contaminated and distorted by bias that the study results lack probative value.

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. Of course, a detailed inspection of those literatures is beyond the scope of the current paper. Another interesting similarity between this literature and those literatures is the conflation of cognitive outcomes (e.g., differences in reaction time to identify an aggressive stimulus) with actual aggressive behavior (e.g., blasting someone with noise). Perhaps the cognitive effects are more plausible and more replicable than the behavioral effects.

REVISIONS: We cite the evidence that brief, minimal social priming manipulations do not influence behavior.

3. Reviewer 1 suggests that our results would benefit from a more thorough theoretical treatment. 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. We have clarified that 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.

We include the dispersion models in a supplementary table.

3. We have redone the moderator analyses as suggested, removing the use of standard errors as predictors and as weights. Only one term reached statistical significance. The effect size in the combined adult/child sample in Anderson, Gentile, and Buckley (2007) was significantly smaller than those in adult-only or child-only samples. Notably, this study also has the largest sample size; adjusting for moderation by standard errors accounts for the influence of adult/child samples.

As now conducted, these moderator analyses are redundant with the moderator analyses reported in Anderson et al. (2010) and will not be included in the current manuscript.

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.” We now support this with appropriate citations.

5. Reviewer 3 points out that the caption to Figure 3 misidentifies an outlier. We’d confused Pearson *r* with Fisher’s *z*. The correct effect size is now listed as *z* = 0.53.

6. Other comments lead to small refinements in language and further support of arguments through citation.