Main points:

**I. Where’s the nuance? What’s the take-home? When is there an effect? When is there not an effect?**

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

There is clearly some manner of bias among experiments of aggressive affect and aggressive behavior. Scientists need to be aware that the evidence for these phenomena is probably less strong than has been previously reported. As to whether there is or is not “an effect” in experiments, we think it is premature to try to come to a conclusion. We report adjusted effect size estimates, but in the face of such a problematic evidence base, it is something like trying to squeeze water out of a rock. This is something for future research to determine.

For us, the take-home is the funnel plots. These are alarming funnel plots, and they 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.

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.

Cross-sectional studies do not seem to suffer from bias, but cannot provide evidence of a causal relationship. 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 badly overestimated, and that the knowledge gained through the use of these experiments may represent little more than overfitting.

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.

In general, we do not think it appropriate to try to make a firm pronouncement on the existence or absence of an effect (or statistical significance as a proxy for such) in experiments. Our different models and procedures yield decision statistics that disagree as to whether there is or is not an effect. In the presence of such pernicious bias and such limited evidence, we think it best to suspend judgment.

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

They don’t. We created a supplementary table conducting naïve fixed-effects meta-analyses and trim-and-fill analyses as Anderson et al. did. Their trim-and-fill analyses were restricted to the “best experiments” and “best partials” datasets. Only a few of the effect sizes we removed were part of the dataset subjected to trim-and-fill. Panee and Ballard (2002) is treated as a best-practices study of aggressive affect and physiological arousal, but a not-best-practices study of aggressive behavior. Graybill et al. is listed as not-best-practices, and so is not considered in their trim-and-fill analyses.

Anderson et al. report trim-and-fill results for best experiments *r+* = .294, with zero imputed studies. Excluding Panee & Ballard, we get naïve FE result *r* = .289, and trim-and-fill imputes 6 studies to the left, reducing the effect size to *r+* = .238. Including Panee & Ballard, we get naïve FE result *r* = .308, a little bit higher than their naïve FE result. Whereas their trim-and-fill imputed zero studies, however, ours imputes 6, reducing the effect size to *r+* = .247. We’re not sure what’s responsible for these differences. Considering how nakedly asymmetrical the funnel plot of best-practices aggressive affect experiment effect sizes is, it seems odd that trim-and-fill would have imputed zero studies. Perhaps there was an error on their part?

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 Graybill et al.

**III. Where’s the theory?**

As with “the effect”, we are very reluctant to make strong pronouncements with regard to theory. Our concern is that the evidence base in Anderson et al. (2010) is too warped, and the meta-analytic adjustments too imperfect, to make a firm decision on whether “the effect” “exists.” We are especially reluctant to conflate statistical significance with existence, given that many of the properties of these meta-analytic adjustments are still unknown and the topic of ongoing study. Still, the present results raise some questions.

It seems uncertain that violent video games create an aggressive affective state, as has been claimed. Of the accumulated funnel plots, that of experimental effects on aggressive affect was the most clearly asymmetrical. 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.

It is hard to say what the present results mean for aggression theory writ large, because the truth may simply be that brief violent game exposure is too weak a manipulation to perturb the mechanisms of aggressive behavior. For example, it is possible that aggression is still trained, activated, and enacted according to scripts learned through observation (Huesmann’s script theory), but that this process takes place over a longer period than a half-hour’s lab experiment. Similarly, it seems trivial to suggest that people are more likely to aggress when they are feeling aggressive and thinking aggressive thoughts (Anderson’s General Aggression Model). It may simply be the case that playing a violent game for 15 minutes does not make people feel aggressive or stimulate the kind of aggressive thoughts that are an antecedent of aggressive behavior. Thus, we are reluctant to claim that these results contest the broader ideas of script theory, social learning theory, the general aggression model, etc., insofar as these theories have found support outside of violent game research.

We do point out that “priming” has historically been a hypothesized mechanism of violent-game effects. That is, it has been argued that playing a violent video game “activates” aggressive thoughts, which then inevitably alter behavior. We are skeptical of this account in much the same way that we are skeptical of other such “social priming” effects. It seems plausible that violent video games may make aggressive concepts more readily accessible. To claim, however, that this accessibility must necessarily represent some aggressive intention, or that it must inevitably have some effect on behavior, is suspect. Against the broader context of personality and cognition, this would seem to be a sneeze in a hurricane. We suggest that although, as you say, “all extant theories predict that stimuli aggressive beget behaviors aggressive,” greater attention should be paid to the magnitude of aggressive stimuli. The same applies to social priming results in general.

It seems necessary to revise these theories in the following ways. First, researchers should consider that small manipulations will likely have very, very small effects. Second, greater attention could be paid to what it means to “activate” a thought – there seems to be considerable ambiguity and conflation around this concept between neuroscience (the excitation of a neuron), cognitive psychology (improvements in the discriminability of a word), and social psychology (a broadly-conceived “thought accessibility”), even though each operates on temporal and physical scales that differ by orders of magnitude. Third, insofar as thoughts may indeed be “activated,” consider that activation of thoughts need not necessarily lead to expression of those thoughts – even in the stochastic, between-groups sense.

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.

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.

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 *z* = 0.53.

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