Main points:

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

Regarding the presence or absence of an effect, we’re not sure just what to think. We feel that trying to make a firm pronouncement on the existence or absence of violent-video-game effects is a bad idea given the limitations of the current evidence. You can’t squeeze water out of a rock.

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. Like Reviewer 4, 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.

In experimental research, we find clear evidence of research bias in estimating the effects of violent games on aggressive affect and aggressive behavior. This is particularly true of the Anderson et al. (2010) best-practices subset, which seems to have preferred the inclusion of statistically-significant results.

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 (200X) 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.

In general, we do not think it appropriate to try to make a firm pronouncement on the existence or absence of an effect (vis a vis statistical significance) 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

Clearly, cross-sectional research has demonstrated evidence of reliably detectable correlations between violent video games and aggressive outcomes. However, these paradigms measure the correlation between the two, and not the effect size of the *causal* influence of violent games on aggressive outcomes. Other studies have attempted to estimate Granger-causal effects through the use of partial correlations (Anderson et al., 2010; Ferguson, 2015; Furuya-Kanamori & Doi, 2016). These partial correlations seem to show significant but small effects. 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.

So to sum it up: Cross-sectional correlations? Certainly yes. Granger causality in correlations? Perhaps, but it’s subtle, and we’re not checking that for bias, because it’s complicated. Causal effects in experiments? Well, it’s hard to say. There might be effects on cognitions, The literature is badly distorted by publication, analytic, and selection bias. Our tools for adjusting for these biases are imperfect. You can’t squeeze water out of a rock. We would prefer to leave that question on the table, preferably for a Registered Replication Report. To try to make a firm pronouncement on the size or statistical significance of the effect (or even what statistical significance means in this context) seems like a good way to look foolish in a few years’ time.

Perhaps some of the heterogeneity in experimental effects on aggressive cognition can be resolved by attention to measures. Arriaga et al. (2008) use an emotion Stroop, assuming that greater aggressive cognition would lead to greater activation of Stroop stimuli’s lexical content, further slowing down emotionally-valent Stoop RTs. They found r = .00 with N = 138. Then again, Cicchirillo & Chory-Assad (2005) also find r = .00 despite using the more-typical word completion task. The three largest effect sizes with N > 40 all use the word completion task. The next largest uses the reading reaction time task.

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

They don’t.

We spoke with another Anderson co-author and she said that there was no mistake in entering Matsuzaki et al. (2004). We note the implausible effect size is an outlier and supply sensitivity analysis, but otherwise include it in our analyses. It is a cross-sectional study, which we note are generally robust and unbiased. Thus, effect sizes from Matsuzaki appear in our

The exclusion of Panee & Ballard (2002) had little influence on the results. It is a high-effect-size, low-sample-size study, and so its inclusion would further indicate funnel-plot asymmetry. It does have very small p-values, though, so it would increase the p-curve and p-uniform estimates.

In this study, the IV is whether the participants were urged to kill all the guards in a video game level vs. urged to complete the level as quickly as possible, possibly by killing the guards. The aggressive-behavior DV is whether they attacked the guards in a subsequent video game level. Anderson et al. coded this as “not-best practices,” whereas we feel it is irrelevant. The aggressive-affect DV is the hostility subscale of the Bell Adjustment inventory. This is not a good hypothesis test because participants in either condition may still have experienced violence while freely playing the video game.

The exclusion of Graybill doesn’t influence the results because Anderson et al. coded it as not-best-practices. We looked for effect sizes that were not manipulation checks: they are entered by Anderson et al. as 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. These are, again, inappropriate.

III. Where’s the theory?

Again, 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.

Nevertheless, we now better explain some of the theory behind how and why there are hypothesized to be violent game effects. Notably, “priming” has historically been one of these; the idea 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 extremely suspect. Against the broader context of personality and cognition, this would seem to be a sneeze in a hurricane.

It also seems uncertain that violent video games create an aggressive affective state, as has been claimed.

It is hard to provide the appropriate context with regard to theories of aggression writ large. Perhaps it is possible that aggression is still trained, activated, and enacted according to scripts learned through observation (Huesmann’s script theory). 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). We are very 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. It is possible that brief violent game exposure simply does not provide a sufficiently powerful manipulation with which to test the predictions of these theories.

It seems necessary to revise these theories in the following ways. First, remember that subtle manipulations will likely have minimal effects. Second, recognize that merely “activating” a thought need not lead to expression of that thought – even in the stochastic, between-groups sense. 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, cognitive psychology, and social psychology, even though each acts on a wildly different temporal and physical scale.

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 separating the wheat from the chaff and in directing 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 even more minimal manipulations in stimulating aggression (e.g., black-color primes, heat-word primes, pictures of guns). We have not been able to find the relevant citations for these 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 that is very different from mere lexical priming. 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.

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, but you get the idea. 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. 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.)

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 (Moreno et al., 2009) using weighted fixed-effects models with multiplicative error terms. These were fit using the standard lm() function. 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 include that 1) PET and PEESE estimate a slightly larger effect of violent games on aggressive behavior in best-practices experiments, *r* = .12 and *r* = .17, 2) the PET estimate of the effects of violent games on aggressive behavior in best-practices experiments is now statistically significant, and 3) cross-sectional correlations with aggressive cognition are estimated as much larger

[MORENO ET AL. 2009 SEEMS TO RECOMMEND PETERS OR D-VAR OVER RE-VAR. THIS SUGGESTS I SHOULD USE MULTIPLICATIVE ERROR INSTEAD OF RANDOM EFFECTS. YUCK!]

[We have redone the moderator analyses as suggested. Meta-regression is PET and PEESE. Weights removed in moderator\_analyses.R]

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

We thank Reviewers 2 and 4 for their kind words.