Evaluating Evidence from Non-Significant Results: A Bayesian Perspective on Violent Games Research

Joseph Hilgard, Jeff Rouder, Christopher R. Engelhardt, Bruce D. Bartholow

Abstract

Despite decades of research, the purported association between violent video games and aggressive remains controversial. One source of controversy stems from questions of experimental control, as advocates of the effect argue that the effect remains when stimuli are matched in all irrelevant dimensions, while detractors of the effect argue that the effect is eliminated under better-matched conditions. Both these arguments require statistical evidence for the null hypothesis, which cannot be provided by the use of null-hypothesis significance testing. To evaluate these claims, we apply a more appropriate Bayesian analysis to measure evidence for or against the null hypothesis. We conclude that small-sample pilot tests cannot rule out substantial confounds. Furthermore, we find that studies that claim to find an absence of violent video game effects vary substantially in the strength of evidence, with some even finding evidence of an effect. We recommend the use of Bayesian analyses, larger sample sizes, and the creation of custom-designed games for experimental research.

Despite more than two decades of research, scientific opinion on whether violent video games cause aggressive outcomes remains divided and the research literature controversial. To date, this relationship has been examined by four different meta-analytic teams, two of which argue that there is a meaningfully large effect (Anderson et al., 2010; Greietemeyer & Mugge, 2014) and two of which that argue there is no meaningful effect (Ferguson; Sherry).

In this debate, a major point of contention has been the degree to which the observed effects instead may be caused by confounding game features other than violent content. When stimuli are appropriately matched on these confounds, it is argued, the effect is eliminated. For example, Adachi and Willoughby (2011) argue that it is competition, not violence, which causes increases in aggressive behavior, and that matching game stimuli on competitive content eliminates the purported effect of violence. Similarly, research by Przybylski et al. (2014) indicates that changes in aggressive affect may be due to difficult, competence-impeding controls, rather than violent content. Finally, research by Elson [et al.] (2014) argues that changes in aggressive behavior are caused by games' differences in pace of action, not violent content.

Each of these above arguments infers that, under certain conditions, the null hypothesis is true. However, this inference cannot be supported through the use of *p*-values and null hypothesis significance testing [NHST] using the nil null hypothesis H_0: $\delta = 0$. This statistical approach can reject the null hypothesis in favor of an alternative hypothesis H_A: $\delta \neq 0$, thereby providing evidence for an effect, but it cannot reject the alternative hypothesis in favor of the null hypothesis. A *p*-value greater than .05 could reflect the truth of the null hypothesis, but it could also represent a true effect studied with insufficient power. The statistical analyses presented by the above studies, then, cannot quantify the accumulated evidence for the null hypothesis, if any.

Arguments for the null hypothesis are also common in this literature, even outside the above studies. A common scenario is pilot testing. The experimenter gathers ratings of stimuli from a (usually small) sample of subjects, hoping to find evidence in favor of the null hypothesis that the two stimuli do not differ on any confounding dimensions. This practice of pilot testing has been deemed a necessary criterion of best-practices studies in some meta-analyses (Anderson et al., 2010), despite the impossibility of concluding in favor of the null hypothesis on the basis of $p > .05$.

In the present manuscript, we reexamine select studies of the relationship between videogame violence and aggressive outcomes in order to better assess the degree of evidence for these null and alternative hypotheses. First we outline an approach for evaluating the matching of stimuli in pilot testing. This approach is then applied to evaluate the results of some previous pilot tests in the research literature. Next, we examine studies which have argued the truth of the null hypothesis of no violent game effect, especially those studies in which improved experimental controls are thought to have eliminated the previously-observed effects of game violence on aggressive behavior.

Providing evidence for the null hypothesis

Imperfect alternatives to nil-hypothesis NHST

Two alternatives to nil-hypothesis NHST come to mind. First, one could perform a null hypothesis test against a second, non-nil null hypothesis. For example, when failing to detect an anticipated effect, one could test against the expected effect size δ with the secondary null hypothesis H_{02}: $\mu_1 - \mu_2 = \delta$. If the study retains H_0 while rejecting H_{02}, it could be argued that the study data are sufficiently unlikely given that the true effect size is δ (e.g., Simonsohn, Simmons, & Nelson, 2014). However, this approach does suffer from the typical NHST problem of dichotomous inferences. Dichotomous NHST procedures cannot discriminate between “no evidence”, “a little evidence” and “a lot of evidence,” instead concluding simply either “there is evidence” or “there is not yet evidence.” The problems of this dichotomization are particularly salient when one considers how slight changes in *p*-value lead to opposite conclusions, such as how the null is rejected at $p = .049$ but the null is retained at $p = .051$. NHST also cannot handle small amounts of evidence well. Given slight evidence, either the null is retained and the slight evidence is mislabeled as “no evidence”, or the null is rejected and the effect size is grossly misestimated.

A second alternative is to instead quantify the effect size and its confidence interval [ESCI]. This does have the advantage relative to NHST of being continuous in quantification. However, ESCI provides neither quantifiable nor inferentially consistent statistics (see Morey, Hoekstra, Rouder, Lee, and Wagenmakers, submitted), and when making inferences using ESCI, researchers seem to mentally convert them to NHST anyway (Hoekstra, Morey, Rouder, & Wagenmakers, 2014). While it is true that values near the ends of the confidence interval are less likely, one cannot know exactly *how much less likely* they are. Similarly, a wide CI indicates that more samples would be necessary to provide a more precise estimate of the effect size, but at what point does the CI become *sufficiently precise* for inference? ESCI is, in our opinion, a useful descriptive tool, but does not permit inferences about the strength of evidence.

Comment [J1]: Maybe cite Cumming or someone who sings praises of ESCI.

Bayesian Statistics

Comment [J2]: Can this section be trimmed?

We propose Bayesian model comparison as the ideal inferential approach when studying and testing the absence of effects. This statistical approach specifies an alternative hypothesis, then compares how probable the data are under the null and alternative hypotheses. When the effect size is near zero, the data are more probable given the null hypothesis than they are given the alternative hypothesis. As the effect size moves away from zero, the data become less probable given the null and more probable given the alternative hypothesis. Increasing sample sizes yield a more precise estimate of the effect size and may exaggerate the difference in probabilities between the two hypotheses. Bayes factor describes the change between beliefs before and after observing data as articulated by Bayes' theorem:

$$Pr(H_0 \mid Data) / Pr(H_1 \mid Data) = Pr(Data \mid H_0) / Pr(Data \mid H_1) * Pr(H_0) / Pr\ (H_1)$$

Comment [J3]: LaTeX?
Maybe a figure / graph?

The ratio of the probability of the data given the two hypotheses is called the Bayes factor and represents the weight of evidence for one hypothesis over the other. This is represented above by the term $Pr(Data \mid H_0) / Pr(Data \mid H_1)$. The Bayes factor is in continuous odds units ranging from 1-to-infinity (indicating perfect evidence for one hypothesis) to infinity-to-1 (indicating perfect evidence for the other hypothesis). A Bayes factor of or near 1-to-1 indicates that the evidence are inconclusive and that either hypothesis predicts the data equally well. Since the emphasis of this manuscript is on providing evidence for the null, we will refer throughout this manuscript to the Bayes factor BF_{01}, the strength of evidence for the null hypothesis over the alternative hypothesis. Thus, a BF_{01} of 2-to-1 favors the null hypothesis, while a BF_{01} of 1-to-2 favors the alternative.

Bayesian statistics describe the change in beliefs as a function of observed evidence. Beliefs before seeing the data are called the "prior beliefs" or "prior odds", and beliefs after seeing the data are called the "posterior beliefs" or "posterior odds". To reach the posterior beliefs, Bayes theorem simply takes the prior beliefs and multiplies them by the Bayes factor. For example, if the null and alternative hypotheses initially seem equally probable (1-to-1 odds), and the Bayes factor indicates 3-to-1 evidence in favor of the null hypothesis, then the null hypothesis is now favored with 3-to-1 odds. If the null hypothesis seems, a priori, highly probable (say, 10-to-1 odds), and the Bayes factor is 2-to-1 in favor of the null, then the null hypothesis is now given 20-to-1 odds. When the data are incapable of discriminating the null from the alternative, the Bayes factor is 1-to-1, and the posterior odds are equal to the prior odds – the data have not changed beliefs. This is a substantial improvement over NHST, in which the same $p > .05$ test statistic could mean either than the null is true or that the data are insufficient, which prevents researchers from increasing their belief in the null hypothesis. This Bayesian approach is similarly an improvement over ESCI in that it describes precisely how much less likely values at the edge of a CI are, whether that constitutes evidence for or against a particular hypothesis, and if so, what quantity of evidence is provided.

Specifying an Alternative Hypothesis.

In order to perform a Bayesian analysis, it is first necessary to specify an alternative hypothesis. This may sound daunting at first, but it is quite possible for anyone who consumes research with some attention to effect sizes. In this alternative hypothesis, we describe the distribution of hypothesized values of the effect size. This is in contrast to traditional analyses, which assume a single true effect size. An alternative hypothesis can be as specific or as diffuse as necessary: "The effect is $\delta = 0.4$ with 95% CI [0.0, 0.8]," or "the effect is greater than zero, with smaller values more likely than larger values," or "there is some effect in some direction" are all feasible alternative hypotheses. One particularly useful alternative hypothesis is the JZS Default Prior (Rouder citation needed), which models the effect size as δ as a Cauchy-distributed variable (think a normal distribution with much fatter tails) with the degree of spread specified by the analyst; this minimally-informative hypothesis can flexibly describe many effects, but respects that small effects are more likely than large effects.

It is also possible to specify and compare more than one alternative hypothesis. This approach can be useful when two competing hypotheses would predict effects of different magnitudes or directions. It is also helpful when assessing the results of a replication: one alternative hypothesis can broadly describe the anticipated effect, while another alternative hypothesis can specifically describe the effect as obtained in previous research. (See Boekel et al. (in press) for an example.) For an accessible introduction to the practice of specifying an alternative hypothesis and appropriate software tools, we suggest the interested reader consult recent work by Dienes (2011; 2014) and by Rouder et al (2012a, 2012b).

Arguing the Null in Pilot Testing of Matched Stimuli

We apply the above approach to interpreting the results of stimulus-matching pilot testing. Suppose we are designing a study to see whether violent content in games influences aggressive behavior. Participants will play one of two games (violent or nonviolent) and then have an opportunity to aggress against a confederate. In order to make a causal statement that the observed effects, if any, are specifically due to violence, it is useful to first make sure that the two games are alike in all dimensions save violence. We run a small pilot study (n = 20), asking each participant to rate each game for violence, difficulty, arousal, and enjoyment. Performing paired-samples t-tests on each outcome, only violence is found to significantly differ, $p < .05$. We might be tempted to conclude, then, that the two games are matched on the other outcomes. However, this conclusion does not follow on the basis of $p > .05$ alone.

In the research literature on violent games, advocates have suggested that this process of matching is one of the criteria that separate “best practices” studies that find larger effects from “not best practices” studies that find smaller effects (Anderson et al., 2010). At the same time, skeptics have suggested that matching games on certain dimensions eliminates the effect of violent games (Adachi & Willoughby, 2011). However, interpretation of these pilot tests has been improper and incoherent. For example, pilot tests in this research domain have sometimes estimated the differences between stimuli as being large, but because the results were not statistically significant, the null hypothesis was considered confirmed. In one particularly remarkable case, post-hoc Bonferroni correction for multiple comparisons was applied to control the Type I error rate across comparisons on 14 dimensions, lowering the critical value of p to .0036 (Arriaga, Esteves, Carneiro, & Monteiro, 2008). Differences as large as r = .53 were observed but not considered statistically significant due to the small sample size and harsh multiple comparison correction. To their credit, the authors acknowledge that the pilot sample was small, but still do not entertain the possibility that the pilot test provided evidence of differences; instead, they conclude that the pilot test indicates that the games are relatively well-matched.

Pilot tests constructed this way are impossible to support through the use of NHST, because they are constructed so that the researcher is on the wrong side of the null hypothesis: trying to demonstrate the truth of the null with a statistical method that can only reject the null. Worse, the more data that is collected, the better the statistical power to detect a confound, and the more likely it becomes that one or more confounds will emerge as significant. This inferential approach, then, will reward researchers for collecting insufficient data and risks failing to detect substantial confounds. Indeed, with a sufficiently small pilot and harsh enough multiple comparison corrections, even large confounds will go undetected.

Bayesian Analysis in Pilot Testing

Bayesian analysis provides a proper approach to testing whether stimuli are matched. To test whether two stimuli are matched, one specifies a null hypothesis of no difference (H_0: $\delta = 0$) and an alternative hypothesis of a moderate difference (e.g., H_A: $\delta \sim$ Cauchy(scale = .5)). If it is unreasonable to expect that the stimuli are perfectly matched, and small differences would be considered acceptable, a null hypothesis of minimal difference can be used instead (e.g., H_0: $\delta = 0$ or $\delta \sim$ Uniform(-.1, .1), see the `nullInterval` argument for the `ttestBF` function in the BayesFactor R package). Stimulus ratings are gathered, and the probabilities of the data given the null hypothesis and given the alternative hypothesis are compared. If the Bayes factor favors the null ($BF_{01} > 1$), the researcher has evidence that the two stimuli do not differ on the particular dimension. If the Bayes factor favors the alternative ($BF_{01} < 1$), this is evidence that the two stimuli do differ. Finally, if the Bayes factor favors neither hypothesis ($BF_{01} \approx 1$), the data are not sufficient to discriminate between the two hypotheses.

This approach rewards researchers for collecting more, rather than less, pilot data. Because Bayes factors are insensitive to stopping rules (Rouder, 2014), the researcher may return to collect additional pilot data if the first wave of collection proves inconclusive. But how much evidence is needed? Recall that posterior beliefs are the product of prior beliefs and the Bayes factor. In the case that two stimuli seem to be obviously matched, it may not be necessary to provide a lot of evidence in a thorough pilot test; in the case that two stimuli would seem to be poorly matched, substantially more thorough pilot testing will be necessary to demonstrate their matchedness. There can be no objective threshold that separates "sufficient evidence" from "insufficient evidence", as prior beliefs are inherently subjective. Thus, to the question "How much evidence do I need?" the answer is simply "Enough to convince your reviewers, readers, critics, and yourself." Rouder, Morey, and Wagenmakers (submitted, p. 12) explain the value of evidence in the absence of a decision rule:

> Finely graded evidence may be thought of as a quantity, say like the weight of some number of bananas. If one has a pound of bananas, there is no reason to make a decision

> whether a pound is a significant weight of bananas. We may all agree that it is what it is, a pound, even though it may have different meanings to differently sized monkeys, say gorillas and spider monkeys. For a pound will satiate a spider monkey but not a gorilla, and so it is with evidence. We may all have our own thresholds but still agree a Bayes factor of 5 is a Bayes factor of 5, and in all cases it is half as much as a Bayes factor of 10 and twice as much as a Bayes factor of 2.5.

We will caution that it can take a lot of data to provide evidence against the existence of very small effects, so it may not be feasible to demonstrate that stimuli are matched to arbitrary precision via pilot testing. Researchers will need to consider the magnitude of potential confounds they intend to account for in pilot testing and balance that against the required sample sizes.

Reanalysis of Select Pilot Tests in Violent Media Research

To assess whether pilot tests have provided convincing evidence of the equivalence of matched game stimuli, we perform a Bayesian reanalysis of previous studies and assess the evidence for the null hypothesis. We use the ttestBF function in the BayesFactor package (Morey et al., 2012) to calculate paired-sample or two-sample Bayesian *t*-tests with scale on effect size set to 0.5 and a null interval over [-0.1, 0.1]. That is, to compare the evidence for or against the null, we compare the null hypothesis H_0: $|\delta| < 0.1$ against the alternative hypothesis H_A: $\delta \sim$ Cauchy(scale = 0.5). This choice of scale is subjective, but appropriate. Effects of violent games are expected to be small (e.g., *r* = .21, or about *d* = 0.43), so confounds should be controlled for on a similarly small scale. Increasing this scale variable will increase evidence for the null, while decreasing this scale variable will decrease the evidence for the null; this is because it is easy to demonstrate that there are not large effects, but difficult to demonstrate that there are not small effects. By entering the sample size and the obtained *t*-value of each test, we calculate a Bayes factor describing the strength of evidence for or against the null. [1]

First, we re-examine pilot data from Arriaga et al. (2008). Results are summarized in Table 1. The pilot test, with its sample of *n* = 20 (within subjects), has not provided strong evidence of matching between stimuli on all dimensions. Bayes factors reveal that there is evidence that some dimensions do not differ, but evidence that other dimensions do. After the pilot test, the readers and researchers are forty times more confident that the two games do not differ in involvement and three times more confident they do not differ in presence, boredom, satisfaction, identification, or excitement. However, they should also be twice as concerned that

[1] While this and other analyses in this section would seem to invite a multiple comparisons problem, we remind that Bayes factor expresses evidence, and that multiple comparisons problems are a matter of interpretation, not evidence. “One should not confuse strength of evidence with the probability of obtaining it (Royall, 1997). Evidence is evidence even if, as one increases the circle of what tests are in the “family”, the probability that some of the evidence will be misleading increases.” (Dienes, 2011, pp 280; an excellent resource on this problem).

the games differ in feelings of competence, and four times as concerned that they differ in difficulty. These conclusions are very different from those of the original authors, who interpret the nonsignificant results of the pilot test as indicating that the games are equivalent on all measures, or at worst, that the results might be merely inconclusive. Given that the two video games, *Unreal Tournament* (a first-person shooter game) and *Motocross Madness* (a racing game), come from very different game genres with very different rules of play, and that the evidence indicates differences between games in competence and difficulty, one might be concerned that the observed effects are due to differences in these confounds rather than the effects of violent game content alone.

Another classic pilot test in this literature is found in Anderson et al., (2004, study 1), in which 120 subjects each played one of 10 games (i.e, n = 12 per cell). The games *Glider Pro* and *Marathon 2* were selected as differing in violent content but having nonsignificant differences in other matching variables. Our reanalysis is summarized in Table 1. Evidence for the null hypothesis is slight, and re-analysis indicates that the games instead may differ in their amounts of action. Because we obtain different p-values than the original authors, it is possible that our re-analysis based on summary statistics is yielding slightly different t-values. For instance, a mean squared error is reported for all cells, rather than per-cell SDs, which may cause us to over-estimate or under-estimate the SD of a particular cell. In any condition, the Bayes factor is not likely to change by much, and at this small sample size per cell, will not strongly favor one hypothesis over the other. Further data collection would be necessary to demonstrate the equivalence of these two games on these dimensions.

Similarly, we re-evaluate the pilot test from Valadez and Ferguson (2010). Three game conditions were compared: a segment from the beginning of the open-world shooter game *Red Dead Redemption* (a control condition, argued to be a nonviolent portion of a violent game), a latter segment from that same game (the active condition, argued to be a violent portion of a violent game), and the soccer game *FIFA* (a second control condition, argued to be a nonviolent game). Only a small sample was collected (cell ns = 15, 10, and 15, respectively), and one-way ANOVAs were conducted to detect variance across conditions in ratings of difficulty, competitiveness, and pace of action. Differences in difficulty and competitiveness were reported as not significant, F(2,40) = 2.36, p > .05 and F(2, 40) = 3.09, p > .05, respectively, while differences in pace of action were significant F(2, 40) = 4.27, p = .02. This last variable was explored through Bonferroni post-hoc analysis, and it was decided that the two control conditions differed from each other but not from the active condition.

We perform all pairwise t-tests, then convert these into Bayes factors. Results are summarized in Table 2. Contrary to the author's conclusions, the results of the pilot test indicate that the games are not well matched. Several Bayes factors strongly favor the alternative hypothesis: the two *Red Dead Redemption* conditions differ in Competitiveness, and the two control conditions differ in all dimensions. Most other comparisons are largely uninformative, as might be expected of the very small sample size. Given our prior beliefs that the early levels of a game are often rather easier than the latter levels, that *Red Dead Redemption* and *FIFA* are very different genres of game, and that the evidence indicates differences between the conditions, we are again not convinced that the stimuli are well-matched. Rather than demonstrate that the stimuli are matched, the pilot test has instead indicated that the games are probably quite different. Even large effect size estimates and modest amounts of evidence can result in nonsignificant *p*-values.

Some pilot studies are more successful in demonstrating invariances. Adachi & Willoughby (2011) report two pilot studies intended to demonstrate that the games used (*Conan*, an action-adventure combat game, and *Fuel*, a racing game) were matched on game characteristics but differed in violence. In the first pilot, $n = 14$ participants played each of two games (within-subjects). This pilot provided modest evidence that the two games did not differ in competition, difficulty, or pace of action, BF_{01}s = 3.36, 3.12, and 2.68 in favor of the null, respectively. The subsequent Study 1 provided further slight evidence that the games did not differ, BF_{01}s = 3.04, 1.07, and 2.24 in favor of the null, respectively. ($BF_{01} = 1.07$ is, of course, hardly any evidence at all.) Considering that the two games were, again, from very different genres of game, this might not be enough evidence to conclude that the games are matched stimuli; however, at least the data did not indicate that the games instead differed. Also, neither this study nor Valadez and Ferguson (2012) tested games for equivalence in frustration, so it is possible that other confounds exist but were not tested.

Summary

Because NHST cannot provide evidence in favor of the null hypothesis, it is inappropriate to argue that two experimental stimuli are matched on the basis of a non-significant test result. Through collection of an arbitrarily small sample size and application of harsh post-hoc corrections for multiple comparisons, almost any difference could be presented as “not statistically significant”. Because of the inferential flaws of this approach and the historically small sample sizes used in previous pilot tests, we would not advocate the use of a pilot test as a best-practice criterion in meta-analyzing previous research literature.

As an alternative to NHST, we advocate the use of Bayesian statistics. Evidence presented this way can favor the null hypothesis of no difference, an alternative hypothesis of a confounding difference, or indicate an absence of evidence for either hypothesis. Researchers are rewarded for more thorough pilot testing by larger Bayes factors. These principles apply also to tests of primary hypotheses, as we explore next.

Arguing the Null in Demonstrating Boundaries of Effects

Interpreting Null Results in the Violent Games Literature

The controversy in this research literature has been caused, in part, by differences in study results across researchers. Some researchers report finding statistically significant effects of game violence, while other researchers report retaining the null hypothesis. In some particularly interesting studies, it is argued that the effect has been eliminated through improved experimental controls. Such research suggests that previous studies have overestimated the effect of violent media, mistaking the effects of confounding game features for the effect of violence. If true, this would indicate that effect size estimates from previous meta-analyses (e.g. $r = .21$, Anderson et al., 2010) are in error. Proposed confounds include competitive gameplay (Adachi & Willoughby, 2011), frustrated needs for competency (Przybylski et al., 2014), or pace of action (Elson, Breuer, Van Looy, Kneer, & Quandt, 2014). Research exploring these confounds has attempted to demonstrate both an effect of the confound as well as an invariance with respect to violent content.

To date, sample sizes in some of these improved-control studies have been small. For example, two experiments are reported by Adachi & Willoughby (2011) with total samples of $n = 40$ and $n = 60$. Other experiments are reported by Ferguson and colleagues (2008), Ferguson and Rueda (2010), and Valadez and Ferguson (2012) with sample sizes of $n = 50$ (at least, for subjects randomly assigned), $n = 77$, and $n = 100$, respectively. Przybylski et al. (2014, Studies 1, 2, and 5) perform three experiments with $n = 100$, $n = 100$, and $n = 109$. Another study is reported by Elson et al. (2014) with a sample size of $n = 80$. Assuming that the true effect to be demonstrated or falsified is the $r = .21$ reported in Anderson et al.'s meta-analysis, these studies would appear to be individually underpowered; sample sizes of 40, 60, 80, and 100 would yield one-tailed test power of 38%, 50%, 60%, and 69%, respectively (but note that for a larger effect, such as the expected effect on aggressive affect, $r = .29$, one-tailed power would be 59%, 75%, 85%, and 91%). An ESCI inspection of these studies (Table 3) indicates that many CIs are quite broad, and that many enclose both $r = 0$ and $r = .21$, suggesting that the data are insufficiently precise to favor one hypothesis over the other. However, we nevertheless would like to understand just how much evidence is in each of these studies so that we can assess the validity of the arguments. Because few of these studies use the same paradigm, and many apply new paradigms argued to have eliminated the effect through innovations in experimental control, we cannot combine and meta-analyze studies for greater power. Thus, these single samples of <80% power each are all the evidence that is available for making an inference.

Because these samples are small and the tests underpowered, failure to reject the null may not provide evidence of the truth of the null. This possibility is sometimes dismissed out of hand by authors. For example, Adachi and Willoughby (2011) argue that sample size is not important, saying that "the effect size for game in the current study was zero (partial $\eta^2 = .000$), and thus increasing the sample size would not have made the effect statistically significant." (pp 266). On the contrary, the effect size is measured with error, especially in small samples; increasing the sample size would not only increase the precision of measurement, but also could cause the estimated effect size to change substantially. A similar argument is advanced by Ferguson et al. (2008) "Although the null hypothesis can not traditionally be accepted as 'true,' [Loftus (1996) presented that] if the 95% confidence interval in group difference scores (e.g., μ1

– μ2) is reasonably small, the null hypothesis can be effectively accepted as true. Similarly, [Cohen (1994) suggested examining the confidence interval around the effect size.] Effect-size confidence intervals that cross zero effect can be reasonably concluded to be 'untrue' and, thus, support the null." This approaches an ESCI understanding of the null, arguing that as more data is collected, larger effect sizes can be excluded as being comparatively unlikely. However, given that the effect size confidence interval in that manuscript extended to values greater than the meta-analytic estimate (95% CI on *r* = [-.26, 30]), it does not appear that the 95% confidence interval is "reasonably small" enough to reject the alternative hypothesis in favor of the null.

Comment [b4]: Can turn the bracketed segments into ellipses.

Comment [J5]: Is this too mean? We've all said stupid things about stats in 2008 and 2011.

There is also the problem of near misses in significance testing. For example, one of the study outcomes in Elson et al. (2014) only barely missed statistical significance, *p* = .073. Considering that the estimated effect size (*r* = .20) closely approximated that reported in meta-analysis (*r* = .21, Anderson et al., 2010), it does not seem appropriate to consider this a refutation of the effect. As the saying goes, "Surely God loves the .06 nearly as much as the .05" (Rosnow & Rosenthal, 1989). Instead, it seems likely that this study provides some evidence for the effect, even if this evidence is not sufficiently strong to be considered "significant" by NHST.

Bayesian Model Comparison and Hypothesis Formulation

To assess the strength of evidence for or against the null hypothesis, we re-evaluate these null findings through Bayesian model comparison. We begin by using each study's reported statistics to calculate a *t*-value, the effect size, and the standard error of the effect size.

Next, we specify two alternative hypotheses. First, the effect could be expected to be small-to-medium in magnitude, and a JZS Default Prior could be used to model this. We will refer to this minimally-informative alternate hypothesis as H_{A1}, the first alternative hypothesis. H_{A1} summarizes this hypothesis's predictions about the effect as a Cauchy distribution centered at 0 with a narrow width.

$$H_{A1}: \delta \sim \text{Cauchy}(\text{scale} = .4)$$

By evaluating the probability of this hypothesis relative to the null hypothesis, we create Bayes factor BF_{01}, the probability ratio of H_0 as compared to H_{A1}. As before, when effect sizes are large and have good precision, the data are increasingly improbable given the null relative to the, and the Bayes factor favors this alternative hypothesis, indicating evidence for an effect of small magnitude and nonspecific direction. When effect sizes are near zero, the data are relatively more probable given the null, and the Bayes factor favors the null over this alternative, indicating evidence for no effect.

We also specify a more precise alternative hypothesis. Previous meta-analysis in this research literature provides a specific estimate of the effect as *r* = .21 [.17, .25] (Anderson et al., 2010), which could serve as our alternative hypothesis.[2] We use the meta-analytic effect size estimate and standard error to derive our second alternative hypothesis, H_{A2}:

$$H_{A2}: \rho \sim \text{Normal(mean=.21, sd=.02)}$$

By again comparing the probability of the data given H_0 against the probability given H_{A2}, we create Bayes factor BF_{02}. BF_{02} gives the measure of evidence for the null hypothesis relative to the meta-analytic expectation of the effect size. (Note that the mean and standard deviation used in H_{A2} will vary depending on the particular outcome tested: aggressive cognition, aggressive behavior, and aggressive affect each have slightly different meta-analytic effect size estimates. Displayed above is H_{A2} for the effect of violent game content on aggressive behavior.) For these normally distributed effects, Bayes factors can be easily calculated with the online calculator provided by Dienes (http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes.htm).

With these Bayes factors, researchers can now evaluate an experiment's results as supporting either H_0 or H_{A2}. If $BF_{02} < 1$, the results replicate and support the meta-analytic findings. If $BF_{02} > 1$, the results provide evidence for the null hypothesis, indicating that the null is more likely than the meta-analytic alternative, given the observed data. Comparisons between H_{A1} and H_0 or H_{A1} and H_{A2} could indicate evidence for an effect of a magnitude or direction not predicted by H_{A2}. This model comparison between the null and meta-analytic alternative is applicable in many research contexts in which researchers explore the mediators, boundaries, or potential confounds associated with a psychological phenomenon.

Reanalysis of Null Findings in VVG Research

We apply this approach to the current literature of studies claimed to have found the boundaries of the effect of violent video games on aggressive behavior. Each study has a confidence interval that overlaps with $r = 0$, which caused researchers to retain the null hypothesis and argue evidence for it. However, how much evidence do they provide for the null, if any?

Findings are summarized in Table 3. We find that, among these null findings, the strength of evidence for the null varies substantially. In studies with small sample sizes (Ferguson et al., Study 1; Adachi & Willoughby, 2011, Study 1 and 2), evidence for the null in each experiment is slight. This indicates that the evidence provided by Adachi and Willoughby does favor the null

[2] There exist other meta-analyses in this literature (Ferguson & Kilburn; Greitemeyer & Mugge; Sherry), but this is the most widely-cited of them. If the researcher is of the opinion that meta-analysis has failed to reveal an effect of violent content on aggressive behavior, he or she can use a JZS Bayes default prior. One could also test against a less specific alternative hypothesis that incorporates the uncertainty about meta-analytic conclusions by expanding the variance around H_{A2}'s effect size (e.g. H_{A2}: r ~ N(mean = .20, sd = .1). In this case, H_{A2} is the hypothesis of the researcher's opponent theory, and the researcher hopes to provide evidence against this theory to demonstrate that it is unlikely.

hypothesis of no effect, but that a third, larger experiment might be conducted before we conclude that there is no effect of violent content on aggressive behavior so long as competitive content is matched. In studies with larger sample sizes (Ivory & Kalyanaraman, 2007; Prybylski et al., 2014, Study 1, 2, and 5; Tear & Nielsen, 2014), evidence for the null is much stronger.

In cases where effect sizes were close to *r* = .21 but the confidence interval failed to exclude zero, we do not interpret the study as disproving H_{A2} in favor of H_0. Bayes factors recognize that *r* = .20 much more closely resembles *r* = .21 than it does *r* = .00. Thus, re-examination of the effect of violent game content on noise intensity in Elson et al. (2014) indicates a moderately informative replication. The non-significant result has been misinterpreted as support for the null when instead support has been found for the alternative.

A similar phenomenon is observed in Valadez & Ferguson (2012). In this study, participants' hostile feelings were measured before and after playing one of three games: a section from the beginning of *Red Dead Redemption*, a latter section of *Red Dead Redemption*, and *FIFA.* Participants played the game for either 15 or 45 minutes. The condition in which the participants played the beginning section of *Red Dead Redemption* was considered a nonviolent control condition, as was *FIFA*. Thus, the latter section of *Red Dead Redemption* was compared to the other two conditions, and with a time X group test statistic of *F*(1, 94) = 3.11, *p* = .09, *r* = .17, the authors argued positive evidence for the null hypothesis. On the contrary, compared to the meta-analytic estimate of the effects of violent games on aggressive affect (*r* = .29, [.25, .34], Anderson et al., 2010), the data slightly support the alternative hypothesis, not the null, at 1-to-1.9 odds.

We offer further re-analysis of this study. It seems unlikely that the early section of *Red Dead Redemption* was truly nonviolent. Inspection of game footage indicates that the player-character is shot in a scripted scene within the first 15 minutes of play (see http://youtu.be/3lAB1JlbVIM?t=5m28s). Thus, we performed the analysis again, this time comparing the two *Red Dead Redemption* conditions against the *FIFA* condition. This yields an effect size of *r* = .22, [.02, .39] with BF_{02} of 1-to-8.54, indicating moderately strong support for the meta-analytic alternative. There is one last wrinkle to this study: a main effect of time was observed such that Social Hostility Scale scores *decreased* from pretest to posttest, *F*(1, 94) = 8.15, *p* = .005, *r* = .277 [.078, .443], BF_{01} 1-to-7.7 in favor of the nonspecific alternative. Thus, while this study provides evidence that violent games increase aggressive affect relative to nonviolent games, it also suggests that this observation is not due to increases in aggressive affect as a result of violent gameplay, but rather, smaller decreases in aggressive affect relative to those caused by nonviolent gameplay. (However, remember also that the conditions do not appear to be well-matched, and so this phenomenon could still be due to the same confounds suspected in other research.) Future research should explore this possibility through application of repeated measures designs, when possible, but must also consider the likely failures of deception involved in repeatedly measuring aggressive outcomes immediately before and after violent gameplay.

Comment [DM6]: An advocate of NHST may argue that NHST supports the alternative hypothesis and, hence, the benefit of Bayesian testing is small.

Comment [J7]: I don't know what to say to that other than that even a broken watch is right twice a day!

In summary, while all nonsignificant findings receive the same decision in NHST, a Bayesian analysis provides a more nuanced perspective. Depending on the strength of evidence in a particular study, we might decide that the results reject the alternative hypothesis, in which case a boundary condition of the effect has been identified; the results support the alternative, in which case a boundary condition has not been identified, and the results seem to replicate the broader phenomenon; or the results are inconclusive, and further research would be necessary to determine whether one has found a boundary condition or not.

Still No Replacement for Data Integrity

We describe above how Elson et al. (2013) seem to have found evidence for the theorized effect despite an original argument for the null based on $p > .05$. In correspondence with these authors, they asked that we consider their criticism that the Competitive Reaction Time Task measure of aggression (citation needed) used is flexibly quantified, potentially allowing researchers to selectively report the quantification with the biggest effect size or the smallest *p*-value (Elson, Mohseni, Breuer, Scharkow, & Quandt, 2014). This criticism still holds for Bayesian analyses, as Bayes factors are still a function of the data, and thus, still sensitive to flexibility in quantification. These researchers demonstrated that the same experiment can yield substantially various effect sizes and p-values depending on which quantification strategy is used. In the same way, the obtained BF_{02} and varies substantially depending on the quantification: if mean intensity is used, BF_{02} favors the alternative, 1-to-5, but if mean duration is used, BF_{20} favors neither hypothesis, 1-to-1. Bayes factors for a default alternative hypothesis (BF_{01}) also vary dramatically by quantification strategy (Table 4). As Elson et al. (2014) had noticed, various quantification strategies yielded effect sizes ranging from $\omega = -.32$ (count of low-volume trials, here reported as negative, as it is in the direction opposite to that hypothesized) to $\omega = .00$ (first-trial volume) to $\omega = .39$ (count of high-volume trials). Similarly, BF_{02} ranged from 1400-to-1 (count of low-volume trials) to 3.52-to-1 (first-trial volume) to 1-to-280 (count of high-volume trials).

Summary

Clearly, $p > .05$ can describe a wide variety of situations, and thus, its inferential value is limited. Among the articles reviewed in this section, $p > .05$ applied to a range of all possible study results: some studies had strong evidence for the null, others had only slight evidence for the null, and still others actually supported the alternative. As in the pilot testing example above, failure to reject the null does not constitute evidence for the null; researchers hoping to retain the null can always manage to do so by collecting small sample sizes. While reviewers are becoming increasingly savvy to this problem, there still remains the issue of quantifying the evidence for or against the null, even in a sufficiently large sample. Thus, we advocate the application of Bayesian model comparison techniques presented by Rouder et al. (2012a, 2012b) and Dienes (2011, 2014). These techniques can be used to perform a full 2x2 ANOVA, finding evidence for effects of certain factors while also finding evidence against effects of other factors, as appropriate.

Note that very few of the studies presented in Table 3 exclude *r* = .21 from their confidence interval. Applying a hypothesis test to see if the effect is significantly smaller than *r* = .21 would simply that the data were incapable of rejecting either hypothesis, even though, as our analyses demonstrate, there is at least some evidence in many of these studies. One could instead attempt to interpret the ESCI, arguing that, because *r* = .21 is nearer the extremes of the interval, perhaps some of these studies provide some evidence for the null. However, in the absence of an explicitly defined alternative hypothesis and a Bayesian analysis, it is not possible to know how much evidence the study provides, or even which hypothesis is supported.

Finally, Bayesian analysis is still a function of the data and cannot address concerns about selective reporting. Bayes factor represents the strength of reported evidence. When evidence is selectively reported according to the hypothesis it supports, Bayes factor will be biased. We urge researchers to pre-register their hypotheses and analytic strategies, including method of CRTT quantification. We further urge researchers to attempt a thorough and systematic validation of the CRTT in an attempt to choose a limited number of methods which clearly measure a limited number of constructs. Like any other statistical analysis, Bayesian model comparison is still subject to the problem of “garbage in, garbage out.”

Summary

Making principled and coherent arguments for the null hypothesis is a crucial part of the scientific process. In violent media research, the null hypothesis is frequently argued, first, in matching stimulus materials in pilot testing, and second, in attempting to demonstrate the boundary conditions or absence of an effect. Despite the importance and frequency of these endeavors, traditional statistical practices cannot support these goals. *P*-values greater than a critical threshold do not have any interpretation as supporting the null hypothesis, as they only indicate an absence of evidence for an effect, not an evidence of absence of an effect.

As an alternative, we suggest previously-presented easy-to-use Bayesian alternatives to *t*-tests and ESCI. These Bayesian alternatives require the specification of a reasonable alternative hypothesis. Once researchers have specified an alternative hypothesis, this hypothesis can feasibly be falsified in favor of the null hypothesis. While specification of an alternative hypothesis may sound daunting, it is quite easy, and numerous resources exist to facilitate and evaluate the choice of an alternative hypothesis (e.g., Dienes, 2011, 2014; Rouder et al., 2012).

Our re-analysis found that research in this area would benefit from larger samples and more finely-graded interpretations of results. Inspected pilot studies were largely uninformative, as sample sizes are too small to provide much evidence. Similarly, studies arguing for the null were found to vary substantially in their sample sizes and the strength of evidence for the null. In two cases, a *p*-value very close to the critical threshold was presented as a disconfirmatory finding; re-evaluation of this report indicates instead modest support for the alternative hypothesis. We applaud and encourage research efforts in this area which strive to test the boundaries and causal substrates of the effects (if any) of violent games on aggressive thoughts, feelings, and behavior. However, it is clear from this review that some arguments would benefit from greater evidence. There is strong evidence that violent game contents do not seem to influence aggressive affect independently of player's experienced competence (Przybylski et al., 2014), but weaker evidence that games matched for competitive content do not influence aggressive behavior (Adachi & Willoughby, 2011). Further direct or conceptual replications may be necessary before the evidence is sufficiently persuasive.

We also recommend the collection of larger samples in experimental research. Effects in this research domain are known to be small. Thus, large samples are necessary to discriminate effectively between the null and alternative hypothesis. It is possible that some effects are too small to be feasibly studied in single-institution experiments. Multi-site investigations and antagonistic collaborations could serve a two-fold benefit by increasing sample sizes and by reducing concerns of idiosyncratic effects of particular research teams.

As a further practical recommendation, we note that it may not be feasible to pilot test and match game stimuli to necessary precision. One could potentially invest many subjects in such a test only to find evidence that the games are not well matched. As an alternative to pilot testing commercially-available games for equivalence, we instead favor the approach of software modification. In this approach, researchers take an existing game and modify it with software tools so that the core game is the same, but the construct of interest varies across conditions. It is not unlike adjusting the parameters of computer task script. Because games developed in this way are more obviously matched, as the unmodified portions of the game's code are identical between versions, it requires less pilot evidence to conclude that they are indeed matched. One such manipulation, which involves identical game files which vary in violent content and in the difficulty of gameplay, has been made publicly available for use on OSF (Hilgard, 2014). However, such homemade game modifications will have their limits. It will be infeasible to make professional-quality game modifications with graphics, gameplay, and acting on par with modern popular video games. While research suggests that graphical fidelity is not an important moderator of game effects (Ivory & Kalyaraman, 2007), it is possible that such homemade games do not capture the full real-world phenomenon of video gaming.

We close with an optimistic thought about how Bayesian analysis might further shape the scientific process. It is well understood that, historically, papers finding significant (p < .05) effects have been easier to publish than papers without significant effects (citation needed). This process is thought to contribute to publication bias, in that journals suppress research which does not find significant effects, and questionable research practices, in that researchers muss with their results until the necessary p < .05 threshold is reached. With Bayes factors, there is no such dichotomization or sufficient threshold; instead, evidence is collected and its strength reported. Acceptance of evidence as a continuous quantity may, we hope, reduce journals' and researchers' preference for results that just pass a statistical threshold. By assessing the finely-quantified weight of evidence for each argument from each experiment, we can reach a greater understanding of what is certain, what is uncertain, where evidence is truly contradictory, and where we may simply have come to a misunderstanding.

This nuance is lost in NHST, which provides only dichotomous accept/reject decisions. It is perhaps this dichotomization of evidence which is, in part, responsible for the heated debate in the violent media literature, as each side may misunderstand their rejections or retentions of the null as decisive evidence for or against the effect.

Comment [J8]: A scrap from the cutting-room floor; I sorta like it but IDK where it fits now.

References

Comment [J9]: Double-check completeness of list

Adachi, P. J. C., & Willoughby, T. (2011). The effect of video game competition and violence on aggressive behavior: Which characteristic has the greatest influence? *Psychology of Violence, 1,* 259-274. doi: 10.1037/a0024908

Arriaga, P., Esteves, F., Carneiro, P., & Monteiro, M. B. (2008). Are the effects of *Unreal* violent video games pronounced when playing with a virtual reality system? *Aggressive Behavior*, *34,* 521-538. DOI: 10.1002/ab.20272

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6,* 274-290. DOI: 10.1177/1745691611406920

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. doi: 10.3389/fpsyg.2014.00781

Elson, M., Breuer, J., Van Looy, J., Kneer, J., & Quandt, T. (2013) Comparing apples and oranges? Evidence for pace of action as a confound in research on digital games and aggression. *Psychology of Popular Media Culture.* Advance online publication. doi: 10.1037/ppm0000010

Elson, M., Mohseni, M. R., Bruer, J., Scharkow, M., & Quandt, T. (2014). Press CRTT to measure aggressive behavior: The unstandardized use of the Competitive Reaction Time Task in aggression research. *Psychological Assessment, 26,* 419-432. doi: 10.1037/a0035569

Ferguson, C. J., & Rueda, S. M. (2010) The Hitman study: Violent video game exposure effects on aggressive behavior, hostile feelings, and depression. *European Psychologist, 15,* 99-108. DOI: 10.1027/1016-9040/a000010

Ferguson, C. J., Rueda, S. M., Cruz, A. M., Ferguson, D. E., Fritz, S., & Smith, S. M. (2008) Violent video games and aggression: Causal relationship or byproduct of family violence and intrinsic violence motivation? *Criminal Justice and Behavior*, *35,* 311-332. DOI: 10.1177/0093854807311719

Giancola, P. R., & Zeichner, A. (1995). Construct validity of a competitive reaction-time aggression paradigm. *Aggressive Behavior, 21,* 199-204. DOI: 10.1002/1098-2337(1995)21:3<199::AID-AB2480210303>3.0.CO;2-Q

Ivory, J.D., & Kalyanaraman, S. (2007) The effects of technological advancement and violent content in video games on players' feelings of presence, involvement, physiological arousal, and aggression. *Journal of Communication, 57,* 532-555, DOI: 10.1111/j.1460-2466.2007.00356.x

Morey, R. D., Hoekstra R., Rouder J. N., Lee M. D., & Wagenmakers E. J. (Submitted). The Fallacy of Placing Confidence in Confidence Intervals. Preprint available at http://pcl.missouri.edu/sites/default/files/Morey.etal_.2014.CI_.pdf

Comment [b10]: Check APA formatting.

Przybylski, A. K., Deci, E. L., Rigby, C. S., & Ryan, R. M. (2014). Competence-impeding electronic games and players' aggressive feelings, thoughts, and behaviors. *Journal of Personality and Social Psychology, 106,* 441-457. DOI: 10.1037/a0034820

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review, 21,* 301-308. DOI: 10.3758/s13423-014-0595-4

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research, 47,* 877-903. DOI:10.1080/00273171.2012.734737

Rouder, J. N., Morey R. D., Speckman P. L., & Province J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56,* 356-374. DOI: 10.1016/j.jmp.2012.08.001

Simonsohn, U., Simmons, J. P., & Nelson, L. D. Anchoring is Not a False-Positive: Maniadis, Tufano, and List's (2014) 'Failure-to-Replicate' is Actually Entirely Consistent with the Original (April 27, 2014). Available at SSRN: http://ssrn.com/abstract=2351926 or http://dx.doi.org/10.2139/ssrn.2351926

Valadez, J. J., & Ferguson, C. J. (2012). Just a game after all: Violent video game exposure and time spent playing effects on hostile feelings, depression, and visuospatial cognition. *Computers in Human Behavior*, *28,* 608-616. DOI: 10.1016/j.chb.2011.11.006

Table 1. Pilot test results from Arriaga et al. (2008) and Anderson et al. (2004). Pilot data is largely agnostic between the null and alternative, and in fact sometimes indicates equally strong evidence of certain confounds. BF_{01} ranges from 1-to-infinity (perfect evidence for alternative) to infinity-to-1 (perfect evidence for null). Contrary to the authors' original conclusions, the pilot test has some evidence the games differ in feelings of competence, and fairly substantial evidence that they differ in difficulty. H_0: $|\delta| < .1$; H_1: $\delta \sim$ Cauchy(scale = .5) and $|\delta| > .1$. All Bayes factors rounded to two significant digits.

Comment [J11]: Should pick a consistent font for the tables.

Outcome	*t*	*p*	*R*	BF_{01}
Arriaga et al., 2008				
Difficulty	2.63	.017	.53	1-to-4.0
Competence	2.27	.035	.47	1-to-2.2
Discomfort	1.67	.110	.37	1.1-to-1
Realism	1.56	.135	.35	1.3-to-1
Frustration	1.32	.201	.30	1.8-to-1
Pleasure	1.29	.214	.29	1.9-to-1
Action	1.24	.229	.28	2.0-to-1
Disorientation	1.14	.267	.26	2.2-to-1
Excitement	0.89	.385	.21	2.9-to-1
Identification	0.86	.398	.20	3.0-to-1
Satisfaction	0.83	.419	.19	3.1-to-1
Boredom	0.79	.437	.18	3.2-to-1
Presence	0.53	.601	.12	3.9-to-1
Involvement	0.48	.634	.11	40.-to-1
Anderson et al., 2004				
Action	2.35	.028	.45	1-to-2.6
Difficulty	1.00	.327	.21	1.7-to-1
Frustration	-0.79	.436	-.17	2.0-to-1
Enjoyment	-0.40	.693	-.08	2.4-to-1
Violence	5.48	<.001	.76	1-to-820

Table 2. Pilot test from Valadez & Ferguson, 2010. Pilot testing suggests that the conditions are different, not equivalent, on ratings. BF_{01} ranges from 1-to-infinity (perfect evidence for alternative) to infinity-to-1 (perfect evidence for null). H_0: $|\delta| < .1$; H_1: $\delta \sim$ Cauchy(scale = .5) and $|\delta| > .1$. All Bayes factors rounded to two significant digits.

	Difficulty		Pace		Competitiveness	
Valadez & Ferguson, 2012	*t*	BF_{01}	*t*	BF_{01}	*t*	BF_{01}
RDR "violent" vs RDR "nonviolent"	1.82	1-to-1.3	1.31	1.3-to-1	3.00	1-to-7.2
RDR "violent" vs FIFA	-1.47	1.1-to-1	-2.00	1-to-1.6	0.05	2.5-to-1
RDR "nonviolent" vs FIFA	-3.45	1-to-16	-3.43	1-to-15	-3.00	1-to-7.2

Table 3. Bayesian re-analysis of select studies claiming to find boundaries of violent game effects on affect, behavior, and cognition. Some studies present only modest evidence against the effect, and several indicate evidence for the effect despite nonsignificant *p*-values. BF$_{01}$ = evidence for H$_0$: $\delta = 0$ compared to H$_{A1}$: $\delta \sim$ Cauchy(scale = 0.4). BF$_{02}$ = evidence for H$_0$: $\rho = 0$ compared to H$_{A2}$: $\rho \sim$ Normal(μ, σ), with μ and σ taken from Anderson et al. (2010). Bayes factors range from 1-to-infinity (perfect evidence for alternative) to infinity-to-1 (perfect evidence for null). All Bayes factors rounded to two significant digits.

Comment [b12]: Formatting this table is proving to be challenging.

Comment [b13]: Mention #groups, study design, etc.

Variable and study	*r*	95% CI	n	BF01	BF02
Aggressive affect					
Anderson et al., 2010, Meta-analysis	.29	[.25, .34]	2513		
Valadez & Ferguson, 2012, interaction effect	.22	[.02, .39]	100	1-to-2.3	1-to-8.5
Przybylski et al., 2014, Study 1	.00	[-.19, .20]	99	3.0-to-1	62-to-1
Przybylski et al., 2014, Study 2	.08	[-.11, .27]	101	2.3-to-1	7.1-to-1
Przybylski et al., 2014, Study 5	.03	[-.16, .22]	109	3.0-to-1	38-to-1
Ivory & Kalyanaraman, 2007	.13	[-.05, .30]	120	1.4-to-1	1.8-to-1
Aggressive Behavior					
Anderson et al., 2010, Meta-analysis	.21	[.17, .25]	1454		
Elson et al., 2014, Noise Intensity	.20	[-.02, .39]	84	1-to-1.3	1-to-5.1
Elson et al., 2014, Noise Duration	.11	[-.11, .31]	84	1.9-to-1	1-to-1.1
Ferguson et al. 2008, Study 1 – Random assignment, Noise Intensity	.02	[-.26, .30]	50	2.3-to-1	2.4-to-1
Ferguson & Rueda, 2010 – Violent vs. nonviolent game	.01	[-.21, .23]	77	2.7-to-1	4.4-to-1
Adachi & Willoughby, 2011b, Experiment 1	.00	[-.30, .30]	42	2.2-to-1	2.4-to-1
Adachi & Willoughby, 2011b, Experiment 2	.03	[-.22, .28]	60	2.4-to-1	2.5-to-1
Tear & Nielsen, 2014, hurting behavior in Tangram Task	.01	[-.17, .19]	120	3.6-to-1	9.0-to-1
Aggressive Congition					
Anderson et al., 2010, Meta-analysis	.22	[.18, .25]	2887		
Ivory & Kalyanaraman, 2007	-.08	[-.25, .11]	120	2.4-to-1	130-to-1

Table 4. Bayes factors for each effect size calculated by Elson et al. (2014), study 2, table 2. Bayes factors vary dramatically by quantification method of the CRTT. BF_{01} = evidence for H_0: $\delta = 0$ compared to H_{A1}: $\delta \sim \text{Cauchy}(\text{scale} = 0.4)$. BF_{02} = evidence for H_0: $\rho = 0$ compared to H_{A2}: $\rho \sim N(.21, .02)$. BFs range from 1-to-infinity (perfect evidence for alternative hypothesis) to infinity-to-1 (perfect evidence for null).

Quantification	*r*	BF01	BF02
Mean volume	.20	1-to-1.2	1-to-4.8
Mean volume after wins	.13	1.6-to-1	1-to-1.6
Mean volume after losses	.22	1-to-1.7	1-to-7.2
Mean duration	.11	2.0-to-1	1-to-1.0
Mean duration after wins	.05	2.6-to-1	2.7-to-1
Mean duration after losses	.14	1.5-to-1	1-to-1.9
Mean volume x duration	.18	1.0-to-1	1-to-3.7
Mean volume x sqrt(duration)	.18	1.0-to-1	1-to-3.6
Mean volume x ln(duration)	.16	1.3-to-1	1-to-2.5
Count high volume settings	.40	1-to-140	1-to-280
Count high duration settings	.05	2.6-to-1	2.8-to-1
First-trial volume	.03	2.7-to-1	3.5-to-1
First-trial duration	.01	2.8-to-1	4.9-to-1
Count low volume settings	-.34	1-to-19	1400-to-1