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

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

Abstract

Researchers in the psychological sciences often find themselves testing for invariances. This is a problem when the most commonly-used form of statistical inference, null-hypothesis significance testing (NHST), can only find evidence for variances, not invariances. Specific examples of this problem are apparent in pilot tests, in which researchers hope to demonstrate that two sets of stimuli do not differ on potential confounds, and disconfirmatory replications, in which experimenters hope to demonstrate that a previous finding was the result of Type I error or a confound later controlled for. We review why NHST cannot describe the strength of evidence and explain accessible and practical Bayesian alternatives. Research concerning the effects of violent games on aggressive behavior is used as an example throughout.

Psychology is teeming with effects. Researchers like effects because they allow us to infer the structure of relationships between stimulus and response, cognition and behavior, personality and affect through manipulation: researchers push on one part of the system, examine another part, and, if pushing on part A influences part B, an association is inferred between parts A and B. In this way, we observe that mere exposure creates positive attitudes (Zajonc citation needed), that the endpoints of a scale influence the mean of the scale (anchoring citation needed), and that media influences behavior (citation needed).

However, among all this study of things that change, there is also a need to understand that which does not change. Surely invariance is as important as variance, if not for the sake of discriminant validity alone. However, many researchers find themselves lacking for want of the appropriate statistical tools to demonstrate evidence of invariance.

At present, the primary form of statistical inference in the social sciences is null-hypothesis significance testing [NHST]. Researchers collect data, compute a test statistic, and compare the value of that test statistic against the hypothetical distribution of all possible test statistics one might expect to see if the null hypothesis were true. When the data are sufficiently unusual given the null hypothesis, the null hypothesis is then rejected in favor of a nonspecific alternative hypothesis. In most applications, this null hypothesis is a nil hypothesis of no effect: the predictor variable is not associated with the dependent variable. Rejection of this null hypothesis is taken as evidence for an effect of the predictor on the outcome.

In the case that the data are not unusual given the null, the null hypothesis is retained; however, this is not the same as the null being concluded or accepted. The truth of the null hypothesis cannot be determined from p > .05; such a result could reflect the truth of the null, or it could just as easily reflect the data’s inability to discriminate between the null and alternative hypotheses (e.g. poor statistical power). This is because significance testing is not consistent (Rouder citation needed). Whereas a true effect measured with increasing sample size leads, in the limit, to p-values that tend towards zero, a null effect measured with increasing sample size does not give p-values tending towards 1. Instead, when the null hypothesis is true, *p*-value is uniformly distributed between 0 and 1. There is no particular *p*-value that indicates the truth of the null.

Nonetheless, one often sees *p* > .05 used as an argument to conclude in favor of a null hypothesis of no difference. One common example is pilot testing of stimuli; the experimenter gathers ratings of stimuli from a (usually small) sample of subjects, hoping to demonstrate evidence in favor of the null hypothesis that the two stimuli do not differ on any confounding dimensions. Another example is null experimental findings, or sometimes “hostile replication” or “destructive testing”; researchers replicate an experiment and find no significant effect, or researchers hope to demonstrate that an experimental phenomenon dissipates when certain confounds are controlled for.

In the present manuscript, we examine the inferential challenges of matching stimuli and destructive testing through the example of the contentious and divided literature on the effects of violent video games. In this field, researchers on both sides of the debate often advance arguments for the null hypothesis. For example, researchers finding evidence of changes in aggressive behavior argue that the stimuli are well-matched and that the effects are not due to confounds between stimuli. On the other hand, other researchers use new paradigms that may be better-controlled, find no significant change in aggressive behavior, and argue that the null is true. While both of these arguments represent meaningful and important scientific conclusions, neither can be supported through the use of a *p*-value.

**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 replicate an anticipated effect, one could test against the expected effect size δ with the null hypothesis H02: µ1 - µ2 = δ. If the study retains H0 while rejecting H02, 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 outcomes. Dichotomous NHST procedures cannot differentiate between “no evidence”, “a little evidence” and “a lot of evidence,” instead concluding either “yes evidence” or “no evidence.” This is troublesome when slight changes in *p*-value lead to opposite conclusions, for example, rejection of the null at *p* = .049 but retention of the null at *p* = .051. [Statistical power problem?]

Second, one might advocate instead quantifying the effect size and its confidence interval [ESCI]. This does have the advantage relative to NHST of being continuous in quantification. However, the interpretation of ESCI statistics is neither quantifiable nor inferentially consistent (see Morey, Hoekstra, Rouder, Lee, and Wagenmakers, submitted), and when making inferences using ESCI, researchers seem to mentally convert them to NHST anyway (citation needed, previous ref?). While it is true that values near the ends of the confidence interval are less likely, the question remains of 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?

**Bayesian statistics**

We propose Bayesian model comparison as the ideal approach when studying and testing invariances. This statistical approach specifies an alternative hypothesis, then compares the probability of the data given the alternative against the probability of the data given the null. This statistic, the Bayes Factor, represents the weight of evidence for one hypothesis over the other. The Bayes Factor is in continuous odds units ranging from 0 (indicating perfect evidence for one hypothesis) to infinity (indicating perfect evidence for the other hypothesis). A Bayes Factor of or near 1 indicates that the evidence are inconclusive, and that either hypothesis predicts the data equally well. This is a substantial improvement over the improper use of NHST, where the same test statistic, *p* > .05, results from either the truth of the null or the insufficiency of data. It is also an improvement over ESCI in that it describes precisely how much less likely values at the edge of a CI are, and whether this constitutes evidence for or against a particular hypothesis.

The first, crucial step of Bayesian analysis is to *propose an alternative hypothesis.* To do this, one specifies a hypothetical distribution of the probable values of the effect size, for example, the effect size δ may be somewhere between 0 and 1, with smaller values more likely than larger values. One may even specify and compare several alternative hypotheses. For example, one could also investigate the possibility that the effect size δ is somewhere between -.5 and .5. This is what distinguishes the Bayesian approach from the frequentist: Frequentist statistics assume a single true effect size δ which is estimated in (hypothetical) repeated experiments, while Bayesian statistics allow a researcher to express beliefs about what are the probable values of the true effect size δ. In the typical approach of NHST, the alternative hypothesis is never specified. It is for this reason that the alternative cannot be falsified in favor of the null. This specification of the alternative may seem like an alarming prospect, but it is quite possible for anyone who reads research articles with some attention to effect sizes. By proposing an alternative hypothesis, the researcher can perform a fair test between the two competing hypotheses. We briefly review methods and software tools for specifying alternative hypotheses and performing Bayesian model comparison.

First, and most generally, one can specify a broad alternative hypothesis centered at zero by using a JZS Default Prior (Rouder, Morey, Speckman, & Province, 2012; Rouder & Morey, 2012). This prior can detect effects of either sign, representing the possible effect size as HA: δ ~ Cauchy(scale = *r*). This Cauchy distribution is centered at zero and has broad tails, creating a general two-tailed hypothesis test. This model recognizes that small effect sizes are more probable than large effect sizes, and that the effect could be in either direction. Note the scale parameter *r* in the prior. The JZS Default Prior can be scaled to test effect sizes of various magnitudes. When scale *r* is large, effects are expected to be larger, and when scale *r* is small, effects are expected to be smaller in magnitude. Software tools exist for applying the JZS default prior in independent-group t-tests, paired-sample t-tests, regression, and ANOVA. This prior is easy to use due to its flexibility and its robust software tools.

Alternatively, one can specify and test a more precise alternative hypothesis not centered at zero using the Bayes calculator developed by Dienes (2011, 2014). For example, one alternative hypothesis might describe the hypothesized effect as half a normal distribution, starting at 0, with standard deviation equal to .5. This hypothesis would represent a one-tailed test where the effect is expected to be about δ = .5, with smaller values again more likely than larger values. In the most specific case, one could test an alternative hypothesis modeling the effect as a normal distribution with a non-zero mean, for example, δ ~ N(mean = .4, sd = .2). Such a hypothesis, then, describes the expected effect size as δ = .4 [0, .8], with values near .4 being most probable.

**Arguing the Null in Pilot Testing of Matched Stimuli**

Suppose that we intend to run 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 stimuli 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 often 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, changing 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 concluding support for the null.

Indeed, pilot tests using NHST in this way are constructed so that the researcher is on the wrong side of the null hypothesis. The researcher hopes not to find a significant effect; however, the more data he or she collects, 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 particular statistical procedure, then, implicitly rewards researchers for collecting insufficient amounts of 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**

Through this approach, researchers could specify an alternative hypothesis that the two stimuli are moderately different on any given dimension. Participants rate the stimuli, and the likelihood of the null and alternative hypotheses are compared given the sample’s estimated effect size and sample size. If the Bayes factor favors the null (BF10 < 1), the researcher has evidence that the two stimuli do not differ on the particular dimension. If the Bayes factor favors the alternative (BF10 > 1), this is evidence that the two stimuli do differ. Finally, if the Bayes factor favors neither hypothesis (BF10 ≈ 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 collected additional pilot data if the first wave of collection proves inconclusive. But how much evidence is necessary? In Bayesian analysis, beliefs after seeing the data are simply the product of the beliefs before seeing the data times the Bayes Factor provided by the data. 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 and informative pilot testing will be necessary to demonstrate their matchedness. There is the heuristic that Bayes factors of less than 3:1 are only ‘anecdotal;’ we flinch at the application of arbitrary thresholds to Bayes factors, but agree that more evidence should be preferred to less evidence.

**Reanalysis of Select Pilot Tests in Violent Media Research**

As an example, we apply this approach to several pilot tests from the violent games literature. We use the Bayes factor calculators provided at pcl.missouri.edu/bayesfactor, using either the paired-sample or two-sample t-test calculators as appropriate with scale on effect size set to .5. That is, to compare the evidence for or against the null, we compare the null hypothesis H0: d = 0 against the alternative hypothesis HA: d ~ Cauchy(x0=0, *r* = .5). By entering the sample size and the obtained t-value of the test, we calculate a Bayes factor describing the strength of evidence for or against the null.

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. Bayes Factors range from indicating evidence of no difference BF01 = 4.30 to evidence of a difference BF01 = 0.30. After the pilot test, the readers and researchers are more confident that the two games do not differ in involvement, presence, boredom, etc., but there is little evidence hatt the games do not differ in realism and discomfort. Moreover, there is some evidence that the games differ in feelings of competence, and some evidence that the games differ in difficulty. These conclusions are very different from those of the original authors, who interpret the results of the pilot test as indicating that the games are equivalent on all measures, or at worst, merely inconclusive. Given that the two video games, *Unreal Tournament* (a first-person competitive shooter game) and *Motocross Madness* (a racing game), come from very different game genres with very different rules of play, this may not be enough evidence to indicate that the stimuli are well-matched, especially since some evidence indicates a difference.

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 latter segment from that same game, and the soccer game *FIFA*. Only a small sample was collected (cell *n*s = 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 target condition.

We perform all pairwise t-tests, then convert these into Bayes Factors.[[1]](#footnote-2) Results are summarized in Table 2. The results of the pilot test provide only slight evidence of invariances: the target condition was similar in competitiveness to the control game FIFA, BF01 = 2.90, but it also was very different in competitiveness to the so-called “nonviolent-in-violent” control condition, BF01 = 0.14. The remaining four comparisons were largely uninformative, not exceeding a BF of 2:1 in favor of either hypothesis. The two control conditions were also observed to be very different from each other on all variables, all BFs > 7:1 in favor of the alternative. 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.

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, BF01s = 3.38, 3.18, and 2.81 in favor of the null, respectively. The subsequent Study 1 provided further slight evidence that the games did not differ, BF01s = 3.11, 1.22, and 2.41 in favor of the null, respectively. (BF01 = 1.22 is inconclusive, yes, but still evidence for the null; just extremely tiny evidence.) 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.

**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 post-hoc corrections for multiple comparison, almost any difference could be presented as “not statistically significant”. We instead advocate the use of Bayesian statistics. Evidence thus collected can favor the null hypothesis of no difference, an alternative hypothesis of a confound, or indicate an absence of evidence for either hypothesis. Researchers are rewarded for more thorough pilot testing by larger Bayes factors for the correct inference.

On a related topic, we turn now to the topic of arguments for the null hypothesis when, historically, measurements have not been sufficiently precise to provide a strong argument for either the null or alternative hypothesis.

**Arguing the Null in Demonstrating Boundaries of Effects**

Invariances are often important to our understanding of the mechanisms which give rise to a psychological phenomenon. Consider a hypothetical phenomenon, Outcome Y, that is possibly caused by Factor A, but also possibly caused by the confounding of Factor B with Factor A. To test this possibility, we conduct an experiment which orthogonally manipulates Factor A and Factor B. If we hope to demonstrate that it is Factor A, but not Factor B, that causes Outcome Y, we must demonstrate both an effect of Factor A and no effect of Factor B. However, this latter invariance cannot be demonstrated through application of NHST.

**Interpreting Null Results in the Violent Games Literature**

As an example, we consider the publication of null results in the research literature on violent video games. At present, the most-cited estimate of the effect size of violent games on aggressive behavior in best-practices lab experiments is *r* = .21 (Anderson et al., 2010). It has often been suggested that this effect is not caused by violent game content itself, but rather caused by confounds such as 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 attempt to demonstrate both an effect of the confound as well as an invariance with respect to violent content.

To date, sample sizes in many of these refutations have been small. For example, two experiments are reported by Adachi & Willoughby (2011) with total samples of N=40 and N=60 and p-values very near 1. 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. 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 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. 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.

These sample sizes are modest and suggest that failure to reject the null hypothesis may not actually provide evidence for the null because the sample lacked sufficient statistical power. 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 η2 = .000), and thus increasing the sample size would not have made the effect statistically significant.” This reasoning is flawed. 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. Researchers cannot say with certainty what would happen if a hypothetical additional sample were collected. 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.

There is also the case of certain 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 is possible 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**

We re-evaluate these null findings through Bayesian model comparison. First, we specify the null hypothesis, H0: *r* = 0. Under this hypothesis, the effect size is exactly zero; as the observed effect size becomes larger, or a nonzero estimated effect size is estimated with greater precision, this hypothesis becomes less likely. When sample sizes are large and the effect size close to zero, this hypothesis becomes very likely.

Next, we will specify two alternative hypotheses. First, the effect could be expected to be small-to-medium in magnitude, again using a Default Prior, as before. We will refer to this default, minimally-informative alternate hypothesis as HA1, the first alternative hypothesis. HA1 summarizes this hypothesis’s predictions about the effect as a Cauchy distribution centered at 0 with a narrow width.

HA1: δ~ Cauchy(scale = .21)

By evaluating the likelihood of this hypothesis relative to the null hypothesis, we create Bayes Factor BF10, the likelihood ratio of HA1 as compared to H0. When effect sizes are large and have good precision, the null hypothesis becomes increasingly unlikely relative to this hypothesis, 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 null hypothesis gains in likelihood, and the Bayes Factor favors the null over this alternative, indicating evidence for no effect.

Again, these Bayes Factors can be easily calculated with the online calculator provided by Rouder (<http://pcl.missouri.edu/bf-two-sample>; Rouder, Speckman, Sun, Morey, & Iverson, 2009) or the R package ‘BayesFactor’ (Morey, Rouder, & Jamil, 2014). Methods also exist for Bayes factors for ANOVA designs (Rouder, Morey, Speckman, & Province, 2012).

However, suppose that the pre-existing literature permits a more specific alternative hypothesis. For example, in the study of violent videogames and aggressive behavior, meta-analysis provides a specific estimate of the effect as *r* = .21 [.17, .25] (Anderson et al., 2010).[[2]](#footnote-3) This can be formulated as a specific alternative hypothesis and also tested. Meta-analysis estimates the effect as having mean .21 and standard error .02, which we summarize as our second alternative hypothesis, HA2:

HA2: r ~ N(mean=.21, sd=.02)

When the estimated effect size is close to this interval, HA2 grows in likelihood relative to the other two hypotheses. Again, this likelihood ratio grows with increasing statistical precision (e.g., data). When the estimated effect size is far from this interval, the likelihood of HA2 decreases. By comparing the likelihood of HA2 against that of H0, we create Bayes Factor BF20. BF20 gives the measure of evidence for the meta-analytic expectation of the effect size relative to the null hypothesis. These 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>) or by changing the value of µ in the BayesFactor package (Morey et al., 2014).

With these Bayes Factors, researchers can now evaluate an experiment’s results as supporting either H0 or HA2. If BF20 > 1, the results replicate and support the meta-analytic findings. If BF20 < 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 HA1 and H0 or HA2 could indicate evidence for an effect of a magnitude not predicted by HA2. 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 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: BF20 ≈ 0.38, or about 2.5:1 odds for the null. This indicates that the evidence provided by Adachi and Willoughby does favor the null, but that a third 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; Tear & Nielsen, 2014), evidence for the null is much stronger: Przybylski et al. find BF20 < 0.17 in each study, or about 6 : 1 odds for the null or greater, whereas Ivory and Kalyanaraman obtain BF20 = 0.012, or about 78 : 1 in favor of the null, and Tear and Nielsen obtain BF20 = .096, or about 10 : 1 in favor of the null.

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 HA2 in favor of H0. 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. indicates a moderately informative replication, BF20 = 5.12. 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 (using the Social Hostility Scale; Anderson, Deuser, & DeNeve, 1995) 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. However, compared to the meta-analytic estimate of the effects of violent games on aggressive affect (r = .29, [.25, .34]), the data slightly support the alternative hypothesis, not the null, BF20 = 1.93. Since it seems unlikely that the early section of *Red Dead Redemption* was truly nonviolent – inspection of game footage indicates that the main character is shot in a cutscene within the first 15 minutes of play, much less the first 45 minutes of play (see <http://youtu.be/3lAB1JlbVIM?t=5m28s>) – 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 BF20 = 8.54, indicating moderately strong support for the 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], BF10 = 4.99. 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 could still be due to the confounds suspected in other research.) Future research should explore this possibility through application of repeated measures designs.

**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 arbitrarily 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 model comparison techniques presented by Rouder et al. (2012) and Dienes (2011, 2014).

[How would the results have looked with ESCI or NHST?] Note that very few of the studies presented in Table 3 exclude *r* = .21 from their confidence interval. Application of this testing procedure would report 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.

**Still No Replacement for Data Integrity**

Earlier in this manuscript, we described 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 CRTT measure 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; 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 BF20 varies substantially depending on the quantification: if mean intensity is used, BF20 = 5.12, a moderately informative replication, but if mean duration is used, BF20 = 1.11, indicating that the data are almost perfectly agnostic between H0 and HA2. Bayes factors for a default alternative hypothesis also vary dramatically by quantification strategy (Table 4). As Elson et al. (2014) had noticed, various quantification strategies yielded effect sizes ranging from ω = .32 (count of low-volume trials) to ω = .00 (first trial volume) to ω = .39 (count of high-volume trials). Similarly, Bayes factors ranged from 10,043 : 1 for the null (count of low-volume trials) to 1,858 : 1 for the alternative (count of high-volume trials).

Bayes Factor reports the strength of reported evidence; drawing inferences from that evidence, however, is still dependent on the overall research context of multiple comparisons and possible selective reporting. To present evidence in the best possible context, 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. Previous research has miscited these quantification strategies as having been validated (citation needed!) when it would be more accurate to say only that these quantification strategies have previously yielded statistically significant results. For example, Carnagey and Anderson (2005) write: “An aggressive-energy score was calculated for each trial by taking the square root of the duration of noise chosen for the opponent and multiplying this value by the intensity of the noise chosen. […] Aggressive energy has been shown to be a valid measure of aggressive behavior (e.g., Baron & Bell, 1975; Bartholow, Anderson, Carnagey, & Benjamin, 2005).” The cited studies are not demonstrations of any form of validity, but rather, studies in which this format of CRTT quantification demonstrated a significant result. This logic is circular: the manipulation has a significant effect on the validated CRTT format, and the CRTT format is valid because the manipulation has a significant effect.

**Summary**

Making principled and coherent arguments for the null hypothesis is a crucial part of the scientific process. In this paper, we outlined two common situations in which researchers argue for the null: first, in matching stimulus materials in pilot testing, and second, in attempting to demonstrate the boundary conditions of an effect. The former is necessary for experimental design and precision, while the latter is an important part of determining the specific causal substrates of psychological phenomena and the discriminant validity of psychological measures. 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, only failing to support the alternative hypothesis to an arbitrary degree.

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).

With regard to the specific research literature on violent video games’ effects on aggressive behavior, we find that researchers often intend to argue for the null, either to demonstrate that two games are equivalent in affective content or that a game has no effect on participants’ behaviors. However, studies arguing for the null 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. Researchers are again encouraged to collect large samples to maximize evidentiary value, whether arguing for or against the null.

Another benefit of analysis with Bayes factors is that evidence is continuously quantified. This continuity allows researchers to understand when a little evidence or a lot of evidence is presented. 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 and sometimes acrimonious 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. The re-analyses presented in this manuscript indicate that the evidence provided by individual experiments is often modest, whether for or against the effect, perhaps in part because the anticipated effect is fairly small in magnitude.

Not only does Bayes Factor alleviate psychology’s longstanding bias against the null, but it reduces the pressure on researchers to reach an arbitrary threshold of evidence by rejecting the null. Nobody wants to conduct a study and find that the results have no evidentiary value. However, when sample sizes are small, as they often are in clinical groups and other hard-to-recruit populations, statistical power is poor, and so finding statistical significance is unlikely. Researchers may find attaining this threshold to be an unattainable standard for evidence. Analysis with Bayes factors, being a continuous form of evidence, allows researchers to state what evidence they have, whether it is a little or a lot. Taking this perspective allows journals to publish according to sample size and the strength of evidence, rather than selecting publications according to whether they happened to attain an arbitrarily small *p*-value.

We urge researchers to adopt Bayesian techniques in pilot testing and hypothesis testing. Tools for these analyses are rapidly increasing in availability and ease of use. Adoption of these methods will allow researchers to understand how much or how little evidence they have, whether arguing for or against the null, thereby alleviating research controversy and more accurately representing research conclusions.

**Other trimmings:**

Bayesian investigation of attrition rates per cell (a la Fischer exact test or chi-square test of independence).

References

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

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). Pilot data is largely agnostic between the null and alternative, and in fact indicates differences between stimuli in difficulty. BF01 ranges from 0 (perfect evidence for alternative) to infinity (perfect evidence for null).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Outcome | t | p | r | BF01 |
| involvement | 0.48 | 0.634 | 0.11 | 4.299245 |
| presence | 0.53 | 0.601 | 0.12 | 3.793212 |
| boredom | 0.79 | 0.437 | 0.18 | 3.258827 |
| satisfaction | 0.83 | 0.419 | 0.19 | 3.167648 |
| identification | 0.86 | 0.398 | 0.2 | 3.098291 |
| excitement | 0.89 | 0.385 | 0.21 | 3.028248 |
| disorientation | 1.14 | 0.267 | 0.26 | 2.435033 |
| action | 1.24 | 0.229 | 0.28 | 2.202403 |
| pleasure | 1.29 | 0.214 | 0.29 | 2.088909 |
| frustration | 1.32 | 0.201 | 0.3 | 2.021931 |
| realism | 1.56 | 0.135 | 0.35 | 1.524227 |
| discomfort | 1.67 | 0.11 | 0.37 | 1.322723 |
| competence | 2.27 | 0.035 | 0.47 | 0.543441 |
| difficulty | 2.63 | 0.017 | 0.53 | 0.295291 |

Table 2. Pilot test from Valadez & Ferguson, 2010. Pilot testing suggests that the conditions are different, not equivalent, on ratings. BF01 ranges from 0 (perfect evidence for alternative) to infinity (perfect evidence for null).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Difficulty | | Pace | | Competitiveness | |
| Valadez & Ferguson | t | BF01 | t | BF01 | t | BF01 |
| RDR "hard" vs RDR "nonviolent" | 1.82 | 0.836568 | 1.31 | 1.447863 | 3 | 0.139724 |
| RDR "hard" vs FIFA | -1.47 | 1.297033 | -2 | 0.671977 | 0.047 | 2.901267 |
| RDR "nonviolent" vs FIFA | -3.45 | 0.061151 | -3.43 | 0.063518 | -3 | 0.139724 |

Table 3. Bayesian re-analysis of select studies claiming to find boundaries of violent game effects on affect, behavior, and cognition. Many studies present only modest evidence, and several indicate evidence for, rather than against, the effect. BF20 ranges from 0 (perfect evidence for null) to infinity (perfect evidence for alternative).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable and study | *r* | 95% CI | n | BF10 | BF20 |
| Aggressive affect |  |  |  |  |  |
| Anderson et al., 2010, Meta-analysis | 0.29 | [.25, .34] | 2513 |  |  |
| Valadez & Ferguson, 2012, interaction effect, | 0.22 | [.02, .39] | 100 |  | 18.84 |
| Time (pre-, post-) X Game (Red Dead Redemption, FIFA) |  |
| Przybylski et al., 2014, Study 1 | 0 | [-.19, .20] | 99 |  | 0.02 |
| Przybylski et al., 2014, Study 2 | 0.08 | [-.11, .27] | 101 |  | 0.17 |
| Ivory & Kalyanaraman, 2007 | 0.13 | [-.05, .30] | 120 |  | 0.61 |
| Aggressive Behavior |  |  |  |  |  |
| Anderson et al., 2010, Meta-analysis | 0.21 | [.17, .25] | 1454 |  |  |
| Elson et al., 2014, Noise Intensity | 0.2 | [-.02, .39] | 84 |  | 5.12 |
| Elson et al., 2014, Noise Duration | 0.11 | [-.11, .31] | 84 |  | 1.11 |
| Ferguson et al. 2008, Study 1 – Random assignment, Noise Intensity | 0.02 | [-.26, .30] | 50 |  | 0.41 |
| Ferguson & Rueda, 2010 – Violent vs. nonviolent game | 0.01 | [-.21, .23] | 77 |  | 0.2 |
| Adachi & Willoughby, 2011b, Experiment 1 | 0 | [-.30, .30] | 42 |  | 0.38 |
| Adachi & Willoughby, 2011b, Experiment 2 | 0.03 | [-.22, .28] | 60 |  | 0.37 |
| Estimated by hand from means & SDs given in personal communication. |  |
| Aggressive Congition |  |  |  |  |  |
| Anderson et al., 2010, Meta-analysis | 0.22 | [.18, .25] | 2887 |  |  |
| Ivory & Kalyanaraman, 2007 | -0.08 | [-.25, .11] | 120 |  | 0.01 |

Table 4. Bayes factors vary dramatically by quantification method of the CRTT. BF10 = evidence for HA1: b ~ Cauchy(0, √2/2) relative to H0. BF20 = evidence for HA2: *r* ~ *N*(.21, .02). BFs range from 0 (perfect evidence for null) to infinity (perfect evidence for alternative).

|  |  |  |
| --- | --- | --- |
|  | BF10 | BF20 |
| Mean volume | 0.940993 | 5.48 |
| Mean volume after wins | 0.429778 | 1.62 |
| Mean volume after losses | 1.357797 | 8.84 |
| Mean duration | 0.344409 | 1.03 |
| Mean duration after wins | 0.247241 | 0.38 |
| Mean duration after losses | 0.485173 | 2.02 |
| Mean volume x duration | 0.759511 | 4.06 |
| Mean volume x sqrt(duration ) | 0.756257 | 4.04 |
| Mean volume x ln(duration) | 0.584183 | 2.74 |
| Sum high volume settings | 167.9638 | 1858.42 |
| Sum high duration settings | 0.245088 | 0.36 |
| First trial volume | 0.235626 | 0.28 |
| First trial duration | 0.228516 | 0.2 |
| Sum low volume settings | 19.53389 | 9.96E-05 |

*Model Comparison.* In the case of more complex study designs, such as 2x2 ANOVA, a variety of models can be created and compared. For example, a researcher might propose up to five models to compare for a 2x2 study design: a null model, a model with a main effect of factor A, a model with a main effect of factor B, a model with main effects of factors A and B, and a full model with main effects of factors A and B as well as an AxB interaction. Each effect βj in the models is distributed according to a specified alternative hypothesis, as in the pilot-testing example above, which can be scaled to expect large or small effects.

Suppose, then, that a researcher hypothesizes that Factor A (e.g. competition) effects the outcome, but Factor B (e.g. violence) does not. The degree to which the model with only a main effect of competition is more likely than the other four models constitutes the evidence for this model. Again, use of the ‘BayesFactor’ package for R allows the specification of alternative hypotheses and comparison of model likelihoods. In this research domain, we recommend specifying an alternative hypothesis with a small scale value *r* on the effect size (e.g. .5 or even .25), as effects of violent media are expected to be small; testing against a larger scale value r (e.g., √2/2 or 1) could overstate the evidence for the null by testing against an alternative hypothesis that includes improbably large effects.

Insert code snippet here.

1. While this 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 CITATION NEEDED) [↑](#footnote-ref-2)
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, or, if testing against a hypothesized increase in aggressive behavior, a uni-directional variant thereof. One could also test against a less specific alternative hypothesis that incorporates the uncertainty about meta-analytic conclusions by expanding the variance around HA2’s effect size (e.g. HA2: r ~ N(mean = .20, sd = .1). [↑](#footnote-ref-3)