A Bayesian Reanalysis of Studies in Violent Media Research

Joseph Hilgard, Christopher R. Engelhardt, Bruce D. Bartholow, and Jeffrey N. Rouder

University of Missouri

Joseph Hilgard

jhilgard@gmail.com

Abstract

Despite decades of research, the purported association between violent video games and aggressive outcomes remains controversial. This controversy stems in part from questions of experimental control. In typical experimental research, aggressive outcomes are measured after participants are randomly assigned to play a violent or nonviolent game, and differences between game groups are thought to represent the causal effect of violent video games. Believers of violent-media effects argue that the violent and nonviolent games are matched on all dimensions besides violence, and so observed effects are due to violent content alone. Meanwhile, skeptics argue that the games are not well-matched, and that when they are, there is no effect of violence. Both sides in this argument require statistical evidence for the null hypothesis: believers need to argue that confounding differences between games are zero, whereas skeptics need to argue that the effects of violent content are zero. However, this evidence cannot be provided by the traditional 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.

A Bayesian Reanalysis of Studies in Violent Media Research

Despite more than two decades of research, the scientific literature on whether violent video games cause aggressive outcomes remains divided and contentious. To date, this relationship has been examined in hundreds of individual studies as well as in aggregate by four different meta-analyses. Even the meta-analyses are divided and contentious—two argue that there is a meaningfully large effect (**??**) and two argue there is no meaningful effect (e.g., **??**). Note here that both positions, that video game violence increases aggression and that video game violence has no effect on aggression, are theoretically important and *a priori* plausible. They both deserve serious consideration on an equal playing field.

A typical experiment in this literature tests for an effect of violence on aggressive outcomes by randomly assigning participants to play a violent or nonviolent video game. After gameplay, an aggressive outcome such as hostile affect, aggressive-word accessibility, or aggressive behavior is collected. Levels of the outcome are compared across groups to estimate an effect size and determine statistical significance of the outcome. In theory, then, assessing the effect of violent video-game content should be straightforward, and there is little reason to expect such controversy.

One reason for the difficulty revolves around experimental control. Violent and nonviolent video games are not typically designed to be exactly alike one another except for violent content. Although the experimenter has experimental control over the video game a participant plays, the experimenter does not have experimental control over the content of the video game. This lack of control generates the possibility that the violent and nonviolent games differ in dimensions besides violent content, and that these differences may constitute confounds which are responsible for observed post-play differences in aggressive outcomes. For example, if the violent game is also more arousing and more frustrating than the nonviolent game, these differences may cause increases in aggressive outcomes, even if violent content does not. Thus, violent media experiments often begin with an attempt to demonstrate that the violent and nonviolent games are as similar as possible on all other dimensions so as to minimize the possibility of confounds and support the argument that the observed differences, if any, are due to violent content alone

Some researchers have agued that the violent-video game effect is not due to violent content but to these other confounding factors. For example, **?** argue that it is competition rather than violence that causes increases in aggressive behavior, and that matching game stimuli on competitive content eliminates the purported effect of violence. Similarly, **?** argue that changes in aggressive effect may be due to difficult, competence-impeding controls rather than violent content. Finally, **?** argue that changes in aggressive behavior are caused by differences in pace of action rather than violent content. Each of these arguments favors the position that under certain circumstances there is no effect of video game violence on aggression.

Although the position that there is no video-game-violence effect is plausible and theoretically important, there is a difficult statistical problem in stating evidence for it. Null-hypothesis significance testing (NHST), the nearly ubiquitous approach for inference in this psychological research, may not be used to state evidence for the null hypothesis that the true effect size is zero. NHST may certainly be used to reject the null hypothesis in favor of an alternative hypothesis, 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 may reflect a true effect size of zero, but, alternatively, it also may reflect insufficient power to detect a true nonzero effect. Therefore, it is unknown whether the previously discussed null findings reflect a true null or a lack of power. Needed is a method for stating positive evidence for the null rather than a lack of evidence for an effect.

The importance of the null hypothesis in building theory has been understated in psychology though there are notable exceptions (**??**). In the current context, whether there is or is not an effect has direct theoretical implications on how media influences affect, cognition, and even behavior. Even proponents of video-game-violence effects have the need to state positive evidence for null hypotheses. In all experiments there is a critical need to match features between violent and nonviolent games so that experimental control is maintained over all game features besides violence. When these features are matched, it is argued that the violent and nonviolent game stimuli differ only in violent content, and so any observed differences in aggressive outcomes are caused by violent content alone and not by other confounding differences. So crucial is this matching process that it has been deemed a necessary criterion of best-practices studies in some meta-analyses (**?**).

In the present manuscript, we examine the strength of evidence for video game violence vs. a fairly-treated no-video-game violence null. To do so, we present Bayesian inference which indeed allows researchers to state positive evidence for either hypothesis as determined by the data. In the next section, we present these Bayesian methods and explain how they can be used not only to find evidence for effects of experimental factors, but also evidence for invariance (i.e., the null hypothesis) in outcomes with respect to experimental factors. Following this, we assess whether violent and nonviolent game stimuli appear to be well-matched by reanalyzing several pilot studies in violent video game research. We then examine the strength of evidence of invariance in those studies reporting no significant effect of violent content. Finally, results are summarized and practical suggestions offered for optimally informative research.

Bayesian Inference

At the heart of being able to state evidence for an effect or for an invariance is Bayesian model comparison. Bayesian model comparison has a long history in statistics and even in psychology. Perhaps the first to suggest the methods we cover was Laplace (1829, republished in 1986), and seminal development occurred through World War II and was presented most comprehensively by **?**. **?** were perhaps the first psychologists to recommend the approach and did so with uncommon gusto in their landmark *Psychological Review* article. The method has gained increasing popularity in statistics and psychology in recent years (**?????**) and is rapidly becoming widespread, especially in cognitive domains. The main hurdle to adoption has often been the difficulty of computation and the unavailability of software (**?**), but these hurdles have been largely removed with Morey and Rouder's (2014) BayesFactor library for the **R** statistics language.

In Bayesian analysis probabilities are used to confer a degree of belief on events, parameters, and even theoretically important positions. Analysts start with stated beliefs, and then update them rationally and optimally using Bayes' rule. For updating beliefs about positions, we use the following form of Bayes' rule:

$$\frac{Pr(H_0|\text{Data})}{Pr(H_1|\text{Data})} = \frac{Pr(\text{Data}|H_0)}{Pr(\text{Data}|H_1)} \times \frac{Pr(H_0)}{Pr(H_1)} \tag{1}$$

It is best to start with the term on the far right, $Pr(H_0)/Pr(H_1)$ which is called the *prior odds*. This term describes the researcher's beliefs about the plausibility of the positions before collecting the data. The term on the left, $Pr(H_0|\text{Data})/Pr(H_1|\text{Data})$, called the *posterior odds*, describes the researchers beliefs after collecting the data. The key question is how did the data affect the beliefs, or, restated, what is the strength of evidence from the data. This evidence is described by the middle term, $Pr(\text{Data}|H_0)/Pr(\text{Data}|H_1)$, which is also called the *Bayes factor*. We will denote the Bayes factor with B, and subscript it to indicate which hypothesis is in the numerator and denominator:

$$B_{01} = \frac{Pr(\text{Data}|H_0)}{Pr(\text{Data}|H_1)} \text{ and } B_{10} = \frac{Pr(\text{Data}|H_1)}{Pr(\text{Data}|H_0)}.$$

Bayes factor values range from 0 to ∞ and describe how much more probable the data are under one position than another. For example, $B_{01} = 10$ means that the null is ten times more probable than the alternative while $B_{01} = .1$ means that the alternative is ten times more probable than the null. Infinite support for the null and alternative are obtained when $B_{01} = \infty$ and $B_{01} = 0$, respectively. A Bayes factor of $B_{01} = B_{10} = 1$ expresses equivalency; the data do not discriminate at all among the positions.

One of the key properties of Bayes factors is that they describe changes in beliefs rather than beliefs themselves. Consequently, two researchers may not agree about the plausibility of positions *a priori*, and, in this case, they will not agree about the posterior plausibility. Nonetheless, they may agree about the Bayes factors, the evidence from data. Therefore, the Bayes factor is not dependent on these prior odds and serves as evidence regardless of beliefs about the initial plausibility of positions. In our view, because Bayes factors describe evidence or change in belief rather than belief itself, it is an ideal statistic for scientific communication. This property contrasts favorably with conventional significance testing, which is about making decisions with long-term error rates controlled rather than about expressing evidence from data.

The remaining task is defining the probability of data under a hypothesis. We describe the simple case where the data are normally distributed and the question is whether the true effect size is zero or nonzero. The generalization to more complex cases is straightforward and will be discussed briefly thereafter. Let δ and $\hat{\delta}$ describe a true or theoretical effect size and an observed effect size, respectively. There are two probabilities that need to be computed, $Pr(\text{Data}|H_0)$ and $Pr(\text{Data}|H_1)$. The former is straightforward. For this simple case, $Pr(\text{Data}|H_0)$ is $Pr(\hat{\delta} \mid \delta = 0)$, which is obtained from the t distribution. Figure 1A shows the hypothesis that $\delta = 0$ as an arrow at zero. Figure 1B shows the probability density under this hypothesis for all values of $\hat{\delta}$ for a sample size of 40 observations divided evenly across two cells. The case for the alternative is more complicated. If the alternative is a single point, say $\delta = .2$, then it is relatively straightforward to compute the probability $Pr(\hat{\delta} \mid \delta = .2)$, which is obtained from a noncentral T distribution. This alternative too is represented as an arrow in Figure 1A and the probability density under this alternative is also shown in Figure 1B. The Bayes factor is simply the ratio of the probabilities. So, for example, if the observed effect size is .4, as shown by the circles in Figure 1B, then the probability density for H_0 is 0.18, the probability density for H_1 is 0.32, and the Bayes factor B_{01}, their ratio, is 1.8.

The specification of a point alternative, though often done in power analyses, strikes us as too constrained. In Bayesian analysis, the analyst can consider a range of alternatives. Figure 1C shows the point null and a distributed alternative. Under this alternative, smaller effects are more weighted than larger ones, and positive effects are as weighted as negative ones. The shown alternative is the default one recommended by Rouder and Morey and colleagues (**????**) as being broadly appropriate for research in psychological sciences. The probability density under this alternative for all values of $\hat{\delta}$ is shown in Fiugre 1D, and the density is more diffuse than that for the null. As before, Bayes factor values are computed as the ratio of these probability densities. Figure 1E shows Bayes factor values for the null vs. the Rouder-Morey default as a function of observed effect size. As can be seen, small observed effect sizes correspond to evidence for the null while larger values correspond to increased evidence for the alternative.

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 **?**, 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 **??** and by **?** and **?**.

Imperfect methods for providing evidence for the null

Two other statistical approaches to providing evidence for the null come to mind, but these are inferior approaches to Bayesian analysis. One option is to perform a significance test against a second null hypothesis of some effect. 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., **?**). However, this approach yields dichotomous inferences, a common problem with NHST. Dichotomous NHST procedures cannot discriminate between no evidence, a little evidence, and a lot of evidence, instead concluding simply either that there is evidence or that 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 at all, 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 **?**), and when making inferences using ESCI, researchers seem to mentally convert them to NHST anyway (**?**). 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 there is no way of knowing at what point the CI is sufficiently precise for inference. ESCI is, in our opinion, a useful descriptive tool, but does not permit inferences about the strength of evidence.

Arguing the Null in Pilot Testing of Matched Stimuli

We apply the Bayesian approach described above to interpreting the results of stimulus-matching pilot testing. As described previously, experiments of violent game effects on aggressive outcomes use pilot testing to check for confounding differences between the violent and nonviolent game stimuli. In order to make a causal statement that the observed effects are specifically due to violence, it is useful to first make sure that the two games are alike in all dimensions save violence, thereby indicating an absence of confounds. Suppose that 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 (**?**). At the same time, skeptics have suggested that matching games on certain dimensions eliminates the effect of violent games (**?**). 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 (**?**). 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.

The null hypothesis in such pilot tests cannot be supported by 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

Because Bayesian analysis can provide positive evidence for the null hypothesis of no effect, it permits fair and appropriate tests of whether stimuli are matched. To test whether two stimuli are matched, one compares a null hypothesis of no difference ($H_0 : \delta = 0$) and an alternative hypothesis of a moderate difference (e.g., $H_A : \delta \sim \text{Cauchy}(\text{scale} = .5)$). If it is unreasonable to expect that the stimuli are perfectly matched, a null hypothesis of minimal difference can be used instead to treat very small differences as practically equivalent to zero (e.g., $H_0 : \delta \sim$ Uniform(-.1, .1), see the nullInterval argument for the ttestBF function in the BayesFactor **R** package).

Having specified these two hypotheses, the researcher conducts a pilot test. In this pilot test, participants rate the stimuli on all relevant dimensions. The differences in ratings are quantified, and the probability of the observed differences given the null hypothesis and given the alternative hypothesis are compared. If the Bayes factor favors the null ($B_{01} > 1$), the researcher has evidence that the two stimuli do not differ on the particular dimension. If the Bayes factor favors the alternative ($B_{01} < 1$), this is evidence that the two stimuli do differ. Finally, if the Bayes factor favors neither hypothesis ($B_{01} \approx 1$), the data are not sufficient to discriminate between the two hypotheses.

Because uninformative data yields an uninformative Bayes factor while stronger data yields stronger Bayes factors, this approach to pilot-testing rewards researchers for collecting more, rather than less, pilot data. Again, this is preferable to an NHST approach, in which the desired $p > .05$ can almost always be obtained by collection of small, uninformative samples. In the case that the resulting Bayes factor is not sufficiently compelling, the researcher may return to collect additional pilot data, as Bayes factors are not biased by conditional stopping rules (**?**). But how large must the Bayes factor be to be sufficiently compelling? 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." **?**, 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.

The magnitude of the Bayes factor is not only influenced by the amount of data collected, but the parameters of the alternative and null hypotheses. When the alternative predicts very small effects, its predictions closely resemble those of the null hypothesis; thus, it can require large amounts of data to achieve the necessary precision to discriminate between the two hypotheses. Therefore, it may not be feasible to demonstrate that stimuli are matched to arbitrary precision via pilot testing. Practical considerations of sample size may limit the degree of precision with which stimuli can be matched.

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. To compare the evidence for or against the null, we compare the null hypothesis of no difference $H_0 : \delta \sim$ Uniform(-0.1, 0.1) against the alternative hypothesis of some difference $H_A : \delta \sim$ Cauchy(scale = 0.5). This choice of scale in the alternative hypothesis is subjective, but appropriate. Effects of violent games are expected to be small (e.g., $\rho = .21$, or about $\delta = 0.43$), so confounds should be examined 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. We use the ttestBF function in the BayesFactor package (**?**) 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). 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.

First, we re-examine pilot data from **?**. 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 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 tested 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, we might not believe that the games are well matched. It is possible, then, that the primary results from this study, in which the violent game was associated with greater aggressive behavior, are not caused by violent content specifically, but may be caused instead by experience of competence or difficulty of gameplay.[1]

Another classic pilot test in this literature is found in **?**, 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 a matched pair differing in violent content but not in other dimensions. Our reanalysis is summarized in Table 1. Evidence for the null hypothesis is slight, and reanalysis indicates that the games instead may differ in amount 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 than the authors' analysis based on the original data. One such source of reanalytic imprecision is that single-cell SDs are not available, and so variance had to be approximated with the reported mean squared error. This 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.

Improper inferences regarding the results of pilot testing are also found among skeptics of violent media effects. We re-evaluate the pilot test from **?**. This study used a three-factor one-way ANOVA design to compare a violent game condition to two

[1] While this and other analyses in this section would seem to involve a problem of multiple comparisons, we remind the reader that Bayes factors express 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." (see Dienes, 2011, pp 280, as an excellent resource on this problem).

non-violent game control conditions. In the violent game condition, participants played a segment from the later stages of the open-world shooter game *Red Dead Redemption.* In one control conditon, participants played a segment from the beginning of *Red Dead Redemption*, argued to contain little or no violence because of the early stage of the game, and in the other control condition, participants played the soccer game *FIFA*, a nonviolent game. Only a small sample was collected (cell ns = 15, 10, and 15, respectively, between-subjects), to rate each game on 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.

To determine the strength of evidence for or against invariance, 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. In particular, Bayes factors indicate evidence that 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 stages of a game are often rather easier than the later stages, 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.

Some pilot studies are more successful in finding evidence of invariance. **?** 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 certain game dimensions but differed in violent content. 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, B_{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, B_{01}s = 3.04, 1.07, and 2.24 in favor of the null, respectively. ($B_{01} = 1.07$ is however no evidence one way or the other.) Whether this is sufficient evidence of matching is a subjective question. Considering that the two games came from very different genres (action-adventure, racing), it might not be sufficient to convince everyone that the games are identical in all ways besides violent content. Still, at least the pilot test did not indicate that the games instead differed. Note also that neither this study nor **?** 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. Non-significant test results can almost always be obtained, even if the null hypothesis is false, through collection of an arbitrarily small sample size and application of harsh post-hoc corrections.

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 (**?**, c.f.). As has been demonstrated above, pilot tests from this literature often provide little-to-no evidence that stimuli are matched, and in fact, often indicate that the two stimuli involve some confounds. Future research studies may be able to use larger pilot studies and provide better evidence of matching, but the evidence in the previous literature is not strong.

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 that indicate stronger evidence. These principles apply also to tests of primary hypotheses, as we explore next.

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. Researchers who retain the null hypothesis often consider such a retention as indicating the absence of a true effect (e.g., **????**). 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 by 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., $r = .21$, **?**) are in error. Proposed confounds include competition (**?**), frustrated needs for competency (**?**), or pace of action (**?**). Research exploring these confounds has found significant effects of the confound but nonsignificant effects of violent content.

These nonsignificant results do not necessarily provide evidence for the null hypothesis. In many of these studies, sample sizes have been small, and so the power to detect the theorized effect of violence has been poor. For example, two experiments are reported by **?** with total samples of $n = 40$ and $n = 60$. Other experiments are reported by **?**, **?**, and **?** with sample sizes of $n = 50$ (at least, for subjects randomly assigned), $n = 77$, and $n = 100$, respectively. **?**, (Studies 1, 2, and 5) perform three experiments with $n = 100$, $n = 100$, and $n = 109$. Another study is reported by **?** with a sample size of $n = 80$. Assuming that the true effect size of violent content on aggressive behavior is $r = .21$ as reported in Anderson et al.'s meta-analysis, these studies are 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.

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, **?** 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 **?** "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., μ_1âĂŞμ_2) 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 âĂŸuntrue' 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.

In some cases, the estimated effect size may be very close to that predicted by meta-analysis but the null hypothesis is retained by hundredths of a p-value. Does such an outcome provide evidence for the null hypothesis? For example, one of the study outcomes reported by **?** only barely missed statistical significance, $p = .073$. Considering that the observed effect size on this measure ($r = .20$) closely approximated that reported in meta-analysis ($r = .21$, **?**), it does not seem appropriate to consider this a refutation of the effect. 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 now describe the models to be compared. First, there is the point-null hypothesis, which describes the true effect size as exactly zero: $H_0 : \delta = 0$. Next, we specify two alternative hypotheses, one representing a broad hypothesis and one representing a very specific hypothesis. First, the broad hypothesis is that the true effect is probably small-to-medium in magnitude. We will refer to this minimally-informative alternate hypothesis as H_{A1}, the first alternative hypothesis, and model it with a minimally-informative JZS Prior. H_{A1} thereby summarizes this hypothesis's predictions about the effect as a Cauchy distribution centered at 0 with a narrow width.

$$H_{A1} : \delta \sim \text{Cauchy}(\text{scale} = .4) \tag{2}$$

By evaluating the probability of this hypothesis relative to the null hypothesis, we create Bayes factor B_{01}, the probability ratio of H_0 as compared to H_{A1}. Second, the more precise hypothesis is that the true effect is equal to that estimated by previous meta-analysis, e.g. $r = .21[.17, .25]$ (**?**). We use the meta-analytic effect size estimate and standard error to derive our second alternative hypothesis,

$$H_{A2} : \rho \sim \text{Normal}(\text{mean} = .21, \text{sd} = .02) \tag{3}$$

By again comparing the probability of the data given H_0 against the probability given H_{A2}, we create Bayes factor B_{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.)

With these Bayes factors, researchers can now evaluate an experiment's results as supporting any of these hypotheses relative to each other. If $B_{02} < 1$, the results replicate and support the meta-analytic findings. If $B_{02} > 1$, 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}. [2] 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 evidence of no effect of violent video games on aggressive behavior. Each study has a confidence interval that overlaps with $r = 0$, which caused researchers to retain and argue for the null hypothesis. Our analysis quantifies the strength of evidence 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

[2] Note that the hypotheses created and tested here are just two of many plausible hypotheses and that the reader may develop and test her own hypothesis as well; for example, a reasonable middle ground might be that there is a nonzero effect of violent video games but it has been overestimated through publication bias, leading to an alternative hypothesis that the effect is distributed as a half-normal centered at zero with standard deviation .23.

(Ferguson et al., 2008, 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 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, aggressive cognition; Przybylski et al., 2014, Study 1, 2, and 5; Tear & Nielsen, 2014), evidence for the null is much stronger. Finally, 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 **?** 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 **?**. In this study, participants' hostile feelings were measured before and after playing one of three games: a late-game section of *Red Dead Redemption* (the active condition), an early-game section of *Red Dead Redemption* (one control condition), and *FIFA* (a second control condition). 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 (pre-, post) X game (active, controls 1 & 2) 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 ($r = .29, [.25, .34]$, **?**) , the data slightly support the alternative hypothesis, not the null, at 1-to-1.9 odds.

There is another issue with this study that requires reanalysis. 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). This control condition therefore may better represent an active violent-game condition. 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 B_{02} of 1-to-8.54, indicating fairly strong support for the meta-analytic alternative. However, this effect must be considered in light of an overall decrease in aggressive affect across all conditions from pretest to posttest, $F(1, 94) = 8.15, p = .005, r = .28[.08, .44], B_{01} =$ 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 that are suspected to cause the effect observed in other research.) Future research could explore this possibility through application of repeated measures designs, but be aware that repeated measurement of aggressive outcomes immediately before and after violent gameplay may alert participants to the research hypotheses and invalidate the study results.

In summary, while all nonsignificant findings receive the same decision in NHST, a Bayesian analysis provides a more nuanced perspective by providing a quantification of continuous amounts of evidence. Attention to the strength of evidence will help researchers to determine whether they have evidence of no effect, evidence of an effect, or inconclusive evidence. This evidence tells researchers whether a research finding has been replicated, an effect has been disconfirmed, or a boundary of the phenomenon has been reached.

Still No Replacement for Data Integrity

We describe above how **?** 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 used is flexibly quantified, potentially allowing researchers to selectively report the quantification with the biggest effect size or the smallest p-value (**?**). In this literature, they argue, this particular measure is quantified in many different ways across studies, suggesting that researchers may attempt several different quantifications until one yields statistical significance. These concerns about flexible analysis apply also to Bayesian analyses, as Bayes factors are still a function of the data and thus still sensitive to flexibility in quantification. **?** 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 B_{02} varies substantially depending on the quantification: if mean intensity is used, B_{02} favors the alternative, 1-to-5, but if mean duration is used, B_{02} favors neither hypothesis, 1-to-1. We examine these fluctuations in Bayes factor across quantification strategy in Table **??**. As **?** 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} ranges 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). To minimize potential flexibility in quantification, we suggest that researchers preregister their analyses, share the raw task data, or provide evidence of the validity of a particular quantification, when possible.

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 and may instead represent insufficient sample size. Even when sample sizes are sufficient, there still remains the issue of quantifying the evidence for or against the null. This can be accomplished by the application of Bayesian model comparison techniques presented by **??** and **??**.

Alternatives to Bayes factor exist, in that researchers can conduct hypothesis tests against an expected effect size or examine the obtained effect size and confidence interval, but these approaches are less informative than Bayesian model comparison. Applying a hypothesis test to see if the effect is significantly smaller than $r = .21$ would simply report 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 , it is not possible to know how much evidence this represents, 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."

Discussion

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 demonstrating the boundary conditions or absence of an effect of violent media. Despite the importance and frequency of these endeavors, traditional statistical practices cannot support these goals. As an alternative, we suggest Bayesian model comparison, which allows for fair and principled tests between the null hypothesis and a reasonable alternative hypothesis.

Our re-analysis found that research in this area would benefit from larger samples and more finely-graded interpretations of results. Inspection of pilot studies found that many pilot studies provided little evidence of matching, and in some cases even provided evidence of confounding differences between game stimuli. In similar fashion, studies in this literature arguing no effect of violent game content were found to vary substantially in 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 (**?**), but weaker evidence that games matched for competitive content do not influence aggressive behavior (**?**). Further direct or conceptual replications may be necessary before the evidence is sufficiently persuasive.

We offer two other practical recommendations to improve pilot-testing and primary tests of hypotheses in this literature. First, 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 (**?**). 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 (**?**), it is possible that such homemade games do not capture the full real-world phenomenon of video gaming.

As a second practical recommendation, we ask that researchers collect larger samples in their experiments. Effects in this research domain are known to be small. Thus, large samples will be 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 could help to increase sample sizes. *Antagonistic* multi-site collaborations could be especially productive, having the additional benefit of reducing concerns of bias within individual research teams.

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 are more likely to be published than are papers without significant effects (**??**). This process is thought to contribute to publication bias, in that only research finding an effect gets published, and questionable research practices, in that researchers muss with their results until the necessary $p < .05$ threshold is reached. Both of these processes will lead to overestimated effect sizes and the propagation of Type I errors. 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 an arbitrary 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.

References

Table 1
Results of pilot tests for equivalence

	t	p	r	BF_{01}
Arriaga et al., 2008				
Difficulty	2.63	.017	.53	1-to-3.98
Competence	2.27	.035	.47	1-to-2.19
Discomfort	1.67	.110	.37	1.13-to-1
Realism	1.56	.135	.35	1.32-to-1
Frustration	1.32	.201	.30	1.80-to-1
Pleasure	1.29	.214	.29	1.87-to-1
Action	1.24	.229	.28	1.99-to-1
Disorientation	1.14	.267	.26	2.24-to-1
Excitement	0.89	.385	.21	2.92-to-1
Identification	0.86	.398	.20	3.00-to-1
Satisfaction	0.83	.419	.19	3.09-to-1
Boredom	0.79	.437	.18	3.20-to-1
Presence	0.53	.601	.12	3.91-to-1
Involvement	0.48	.634	.11	40.4-to-1
Anderson et al., 2004				
Action	2.35	.028	.45	1-to-2.61
Difficulty	1.00	.327	.21	1.71-to-1
Frustration	-0.79	.436	-.17	1.98-to-1
Enjoyment	-0.40	.693	-.08	2.39-to-1
Violence	5.48	$< .001$	.76	1-to-818

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

Table 2
Results of Pilot Tests from Valadez and Ferguson (2012).

Table 3
Bayesian re-analysis of claimed null results.

	r	95% CI	n	BF_{01}	BF_{02}
		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	1.44-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	.02	[-.26, .30]	50	2.3-to-1	2.4-to-1
Ferguson & Rueda, 2010	.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	.01	[-.17, .19]	120	3.6-to-1	9.0-to-1
		Aggressive Cognition			
Anderson et al., 2010, Meta-analysis	.22	[.18, .25]	2887	-	-
Ivory & Kalyanaraman, 2007	-.08	[-.25, .11]	120	2.4-to-1	130-to-1

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 $HA_1 : \delta \sim$ 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 **?**. Bayes factors range from 1-to-∞ (perfect evidence for alternative) to ∞-to-1 (perfect evidence for null). All Bayes factors rounded to two significant digits.

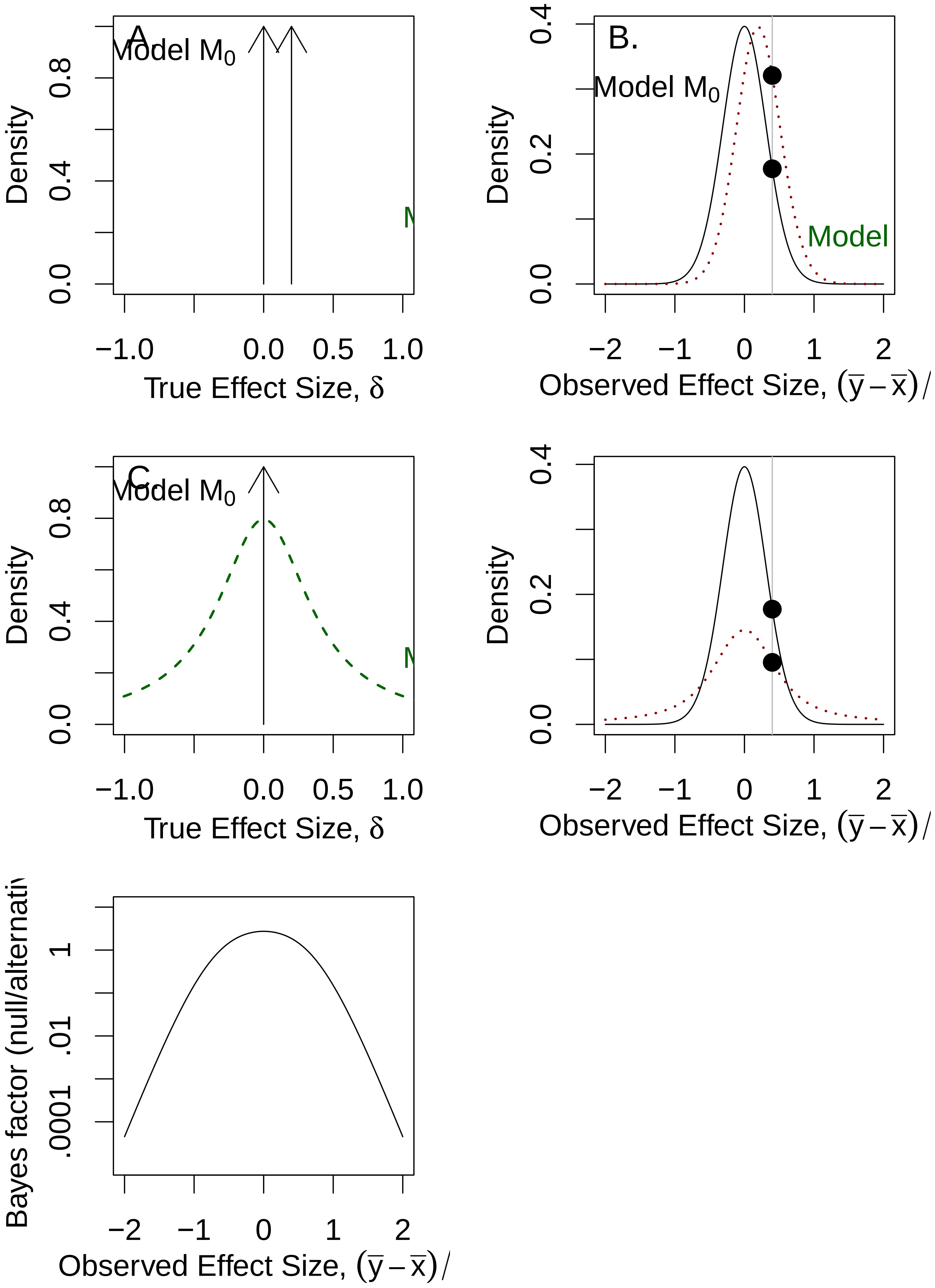

Figure 1. Bayesian model comparison.