International Journal of Law and Psychiatry 56 (2018) 35–43

Contents lists available at ScienceDirect

International Journal of Law and Psychiatry

The problem of false positives and false negatives in violent video game experiments

Christopher J. Ferguson

Department of Psychology, Stetson University, 421 N. Woodland Blvd., DeLand, FL, 32729, United States

ARTICLE INFO

Article history:
Received 4 April 2017
Received in revised form 20 August 2017
Accepted 1 November 2017
Available online xxxx

Keywords:
Video games
Violence
Aggression
Prosocial behaviors
Null results

ABSTRACT

The problem of false positives and negatives has received considerable attention in behavioral research in recent years. The current paper uses video game violence research as an example of how such issues may develop in a field. Despite decades of research, evidence on whether violent video games (VVGs) contribute to aggression in players has remained mixed. Concerns have been raised in recent years that experiments regarding VVGs may suffer from both "false positives" and "false negatives." The current paper examines this issue in three sets of video game experiments, two sets of video game experiments on aggression and prosocial behaviors identified in meta-analysis, and a third group of recent null studies. Results indicated that studies of VVGs and aggression appear to be particularly prone to false positive results. Studies of VVGs and prosocial behavior, by contrast are heterogeneous and did not demonstrate any indication of false positive results. However, their heterogeneous nature made it difficult to base solid conclusions on them. By contrast, evidence for false negatives in null studies was limited, and little evidence emerged that null studies lacked power in comparison those highlighted in past meta-analyses as evidence for effects. These results are considered in light of issues related to false positives and negatives in behavioral science more broadly.

1. Introduction

In recent years, an increased amount of attention has been devoted to the potential that much of what we "know" about behavior may, in fact, be distorted by a publication culture which promotes "statistically significant" results as the expense of null results (Ioannidis, 2005; Simonsohn, Nelson, & Simmons, 2014). Such "false positive" results may be particularly likely in fields which are "hot", which relate to controversial issues of interest to the general public, or which are headline-ready counterintuitive results like to garner considerable attention. For instance, recent years have seen often acrimonious controversies over social priming, a field once considered almost definitively *true* (Bargh & Chartrand, 1999) but now at the center of a replication crisis (e.g. Doyen, Klein, Pichon, & Cleeremans, 2012; Pashler, Coburn, & Harris, 2012). Understanding the mechanisms behind how false positive results are produced in social science research, how they relate to larger social beliefs and pressures, and how scientific culture can either foster or limit them, can be instructive in improving the process of science.

In the current paper, issues related to both false positive and false negatives are considered with the example of video game violence research. Video game violence research exists at the margins of ongoing social concerns about such games, an overlap between science and moral concerns that is ripe for potential problems as described by Ioannidis (2005). Three decades of research (e.g. Dominick, 1984; Graybill, Kirsch, & Esselman, 1985) have been put into examining whether violent video games (VVGs) contribute to player aggression in a meaningful way. Despite that there are now between one to two hundred studies on this topic, little consensus has emerged within the scholarly community about potential effects (see, for example, Consortium of Scholars, 2013). In part this is because studies continue to be published that both do (e.g. Greitemeyer, Traut-Mattausch, & Osswald, 2012; Vieira, 2014) and do not (e.g. Breuer et al., 2015; Charles, Baker, Hartman, Easton, & Kretzberger, 2013) support the view that VVGs contribute to aggression among players. Understanding structural issues that may have limited objective data communication in this field can be illustrative for problems facing social science across similar disciplines with heavy overlap with societal moral debates (e.g. spanking effects, stereotype threat, gender differences, etc.).

One issue to emerge in the larger social science literature, and indeed human sciences including all of psychology, psychiatry and medicine, is the potential for "false positive" results (e.g. Ioannidis, 2012; Pashler & Harris, 2012). False positives occur when researchers reject the null hypothesis for a particular study, despite that the observed effect is the product of chance, sampling error, methodological error, or questionable researcher decisions rather than a "true" effect in the population. Within the field of video game studies the problem of questionable researcher practices (QRPs) which can increase the potential for false positive results has already been identified both for VVGs (Ferguson, 2013) as well as for potential positive effects of

E-mail address: CJFerguson1111@aol.com.

https://doi.org/10.1016/j.ijlp.2017.11.001
0160-2527/

"action" games (Boot, Blakely, & Simons, 2011) which typically happen to also be violent games. The problem of false positives may be particularly likely in a field which is at the center of public and political attention in which politicians or activists are demanding research results to support pre-existing societal concerns (see Griffiths, 2015).

Fortunately, a variety of tools have been developed to test for false positives. False positive results can sometimes be identified through a particular pattern of low power studies with results for statistical significance obtained in higher proportions that would be unexpected given observed power. Given the high standard error of the effect sizes for smaller studies, there is a higher probability in observing small studies with extreme effect sizes, whereas those with extremely low effect sizes will be trimmed away by publication bias. This results in a pattern of effect sizes in which more significant effect sizes are observed than expected given the observed power of the studies (Ioannidis & Trikalinos, 2007). This can be tested through the employment of publication bias analyses (Ferguson & Brannick, 2012), p-curve analysis (Simonsohn et al., 2014), tests for unusual proportions of significant findings given observed power (Ioannidis & Trikalinos, 2007) or through examining the replication probability of a group of studies (Schimmack, 2014). Employing such tools can help alert scholars in a field if their results have been *too good to be true* (Schimmack, 2012) and that they may wish to increase the robustness of their study designs and increase power.

False positives are not the only potential issue for research on VVGs however. Particularly when experimental sample sizes tend to be smaller there is a potential for some studies to report non-significant results when a "real" effect, in fact, exists. This would be a phenomenon of *false negatives* (i.e. Type II error). Just as with the potential for false positives, there are options for examining the potential for false-negatives. One option would be to examine studies with null results using Bayesian statistics which can give a better accounting of the degree to which such studies truly are supportive of the null hypothesis than is typically possible under traditional null-hypothesis significance testing (NHST). Bayesian statistics provide a ratio of probabilities for a dataset under two sets of hypotheses, one of which can be the null hypothesis. Thus, Bayesian statistics allow for a more careful examination of relative support for two potential theoretical models, potentially offering support for null hypotheses.

These issues of false positives and false negatives are examined in three separate sets of analyses, one on a sample of studies of VVGs and aggression identified as "best practices" by a recent meta-analysis (Anderson et al., 2010), a second set of studies on VVGs and prosocial behavior provided by a second recent meta-analysis (e.g. Greitemeyer & Mügge, 2014) as well as a series of recent studies with null results. The authors of the two meta-analyses argued their results supported VVG effects and, as such, these sets of studies will be examined for false positives. By contrast, the studies with null results will be examined for false negatives.

2. Study 1

The first study in this series seeks to examine a set of experimental studies identified as "best practices" (i.e. those studies methodologically best suited to examine hypothesized links between violent games and aggression) by the meta-analysis of Anderson et al. (2010). The purpose of this first study is to examine the fragility of this group of studies to publication bias effects. Furthermore, the R-index (Schimmack, 2014) will be employed to examine the replicability of the studies included as "best practices" in order to examine for the potential that this group of studies may have higher than expected rates of positive results given their observed power.

2.1. Included studies

The list of included studies (k = 27 effect sizes from 22 papers) are provided in Appendix A. These papers were those specifically identified as "best practices" by Anderson et al. (2010) not all studies conducted in the field. Several of the "best practices" studies were from difficult to find sources in Japanese conferences or Japanese language journals, although authors of papers were emailed for copies if they were difficult to find through traditional means. All such studies were located.

2.2. Analyses

All studies were analyzed using both basic correlations between sample size and effect size as well as the more sophisticated Tandem Procedure (Ferguson & Brannick, 2012) for publication bias. The Tandem Procedure is a conservative approach to examining for publication bias by combining several existing publication bias indices (e.g. Egger's Regression, Trim and Fill, etc.) into a decision mechanism. This approach helps reduce the Type I error issues common to stand alone publication bias measures. Specifically, the Tandem's Procedure's decision matrix involves looking for concordance between several issues. First, if the number of studies needed, according to Orwin's FSN to reduce the effect size to a trivial level (typically $r = 0.10$ or lower, although 0.20 also may be used for a higher threshold for practical significance) is equal to or greater than the number of studies observed, this may indicate that the field is exceptionally fragile to publication bias and spurious results. Second, either the rank order correlation or Egger's regression indicates a significant negative correlation between sample size and effect size. Third, Trim and Fill results indicate required adjustment for publication bias. This decision tree approach was developed to reduce Type I error (spurious identification of publication bias). However, it is important to note that, being conservative, the Tandem Procedure may underestimate the true degree of publication bias, particularly bias due to issues unrelated to sample size.

Correlations between sample size and effect size are diagnostic of publication bias because of peculiarities in null hypothesis significance testing (Kühberger, Fritz, & Scherndl, 2014). Specifically, small samples have larger standard error of the effect sizes, producing more extreme effect sizes than larger samples. At the same time, more extreme effect sizes are required for statistical significance with smaller samples. Larger samples have less standard error of the effect sizes, and also do not require large effect sizes to attain statistical significance. Thus, in the presence of publication bias, effect sizes for smaller samples will be larger than for larger samples, given achievement of $p = 0.05$ as a criterion for publication. This creates the negative correlation between sample size and effect size. In the absence of publication bias, no negative correlation should be observed.

Further, studies were analyzed using the R-index. The R-index examines the percentage of studies which achieve statistical significance in contrast to their median observed power. If the proportion of statistically significant studies exceeds those expected given median observed power, this can be an indication that statistically significant results are being selectively reported. As noted by Schimmack (2014, p. 18): "R-index = Percentage of Significant Results − Median (Estimated Power)." R-index calculates observed power and observed *p*-values for individual studies, then calculates median observed power and compares this to the proportion of statistically significant studies. The R-index provides both an inflation rate as an estimate of the proportion of unexpected significant findings given observed power, and an R-index value, which can be considered an estimate of true (rather than observed) median power. Generally speaking, lower R-index values are indicative of greater difficulties with the replicability of a set of studies.

Lastly, p-curve analyses were conducted. Publication bias can occur at multiple levels and for multiple reasons. For instance, publication bias can occur at the level of journals, wherein non-significant results are declined for publication in greater proportions than significant results. Publication bias can also occur at the level of the author, wherein authors either do not submit non-significant results for publication or look for ways to statistically reanalyze their data to convert non-significant results to significant (i.e QRPs). This type of

behavior, *p-hacking* often produces results that are just over the threshold of $p = 0.05$, thus causing a cluster of barely significant results, when *p*-values ought to demonstrate a wider range. This clustering of p-values around 0.05 can be tested with a procedure called p-curve analysis (Simonsohn et al., 2014). Although typically used for clusters of studies from a large paper, or clusters of studies in a journal, examining the p-curve of the "best practices" studies can provide information about whether publication bias is occurring specifically due to p-hacking or potentially due to other factors.

2.3. Results

In the "best practices" analysis of experimental studies of video game violence on aggressive behavior, reported effect size correlates with sample size at $r = -0.503$ ($p = 0.007$), indicating likely publication bias (the negative correlation is - 0.575 when just published Western samples were considered). Results for the Tandem Procedure also indicate clear publication bias with both the rank correlation (tau = 0.436, $z = 3.19$, $p < 0.001$) and Egger's regression ($t(25) = 4.41$, $p < 0.001$) tests for publication bias significant and with the trim and fill, which estimates the likely prevalence of missing studies with null results suggesting that 10 such studies were missing. The significant correlations for the rank correlation and Egger's regression tests indicate that sample size and effect size are inversely correlated, a typical pattern for publication bias as discussed earlier in the paper. Trim and fill informs that the pattern of studies suggest a high probability of the existence of approximately 10 studies with null results that were not reported or included in the "best practices" analysis of Anderson et al. (2010). It is worth noting that any sample of "best practice" studies are, by definition, non-representative. However, it is possible that such studies might truly be selected for methodological strength, in which case publication bias would not particularly be expected, or that conscious or unconscious biases may result in *myside bias* (see Stanovich, West, & Toplak, 2013) wherein studies are valued higher if they produce statistically significant effects. In the latter case, positive publication bias findings are expected.

Regarding the R-index, the included studies were first analyzed using the assumption of typical two-tailed testing with alpha $p = 0.05$ for statistical significance. Observed power of the studies was, on average, 0.547, with but with the rate of having achieved statistical significance of 0.630, indicating an inflation rate of approximately 8.3%, with an R-index of 0.464. Although potentially inflated, the rate by which individuals achieved statistical significance is nonetheless lower than that suggested by Anderson et al. (2010) who concluded that there is overall consistency in violent video game experiments regarding aggression. When a less stringent criteria of $p < 0.10$ was used the rate of having achieved statistical significance jumped to 0.852, despite observed power of only 0.650, suggesting an inflation rate of 20%. R-index was 0.447.

Maintaining the less stringent $p < 0.10$ criteria, the more obscure Japanese language studies were then removed from the R-index, in order to examine issues specifically within high profile journals. Success rate jumped further to 0.909, despite observed power of only 0.690, an inflation rate of 22% with R-index of 0.470. These results suggest that pressure to obtain statistical significance is particularly high in mainstream journals.

When only studies with larger samples (100+) were considered, the picture did not radically change. Here, the correlation between sample size and effect size was actually higher ($r = -0.83$). Success rate was 1.00 (100%), despite observed power of 0.730, indicating a high inflation rate of 27% with an R-index of 0.461. As indicated by Schimmack (2014) R-index values in this range are consistent with a set of studies with relatively low true power for which unsuccessful replications are not being reported.

p-Curve analyses were conducted using the ShinyApps program (http://shinyapps.org/apps/p-checker/). *p*-Values used were observed *p*-values calculated by the R-index. Results for p-hacking were non-significant, indicating that authors converting non-significant results to significant was not the primary driving force of publication bias. Rather, publication bias may be occurring at the level of study selection and publication.

2.4. Discussion

Results of study 1 are consistent with a field of mainly underpowered studies that are achieving statistically significant results to a greater degree than is probable given their observed power. Issues of publication bias were clearly evident for this group of studies, and results from the R-index suggest that results from this pool of studies are improbable given the observed power of these studies. Thus, evidence suggests that this pool of studies may not accurately represent the actual effect size for the link between violent video games and aggression.

p-Curve analyses were non-significant, indicating that conversion of non-significant results to significant through QRPs was not the primary driving force of publication bias in this realm. When publication bias is indicated, but p-hacking is not, the most likely explanation is that publication bias is occurring at the level of study selection. Typically this would indicate journal-level publication bias, although for a meta-analysis, study selection issues may occur at the level of "best practices" nomination.

It is worth noting that p-curve analyses do not necessarily rule out QRPs, only those that "nudge" results over the $p = 0.05$ threshold. However, as noted by Schimmack (2014), many QRPs capitalize on chance, producing *p*-values in a full range, not just those clustered at $p = 0.05$. p-Curve also tends to have less power for heterogeneous sets of studies such as those here. Thus, the absence of findings from p-curve do not rule out the potential for QRPs having occurred in some studies of video game violence.

3. Study 2

Study 1 examined for issues of publication bias and replicability issues in a sample of studies identified as "best practices" in a previous meta-analysis. Study 2 seeks to examine similar outcomes with a sample of studies examining the influence of violent video games on "prosocial" behavior (Greitemeyer & Mügge, 2014). As with study 1, the sample of included papers in study 2 will be analyzed for both publication bias and replicability issues with the R-index.

3.1. Included studies

The list of included studies are provided in Appendix A. The effect sizes of included studies ranged remarkably in this dataset from $d = -0.80$ (Bösche, 2010) through $d = 0.89$ (Greitemeyer et al., 2012). Thus, considerable heterogeneity in study results is evident for this meta-analysis.

3.2. Results

As with study 1, included papers were analyzed using the Tandem Procedure and the R-index. Thus it was difficult to identify a pattern in the results in the studies and with such heterogeneity ($I^2 = 58.42$, $\tau = 0.118$) that it is unlikely that the meta-analytic effect size is meaningful without careful moderator analyses. With such heterogeneity in effect sizes going both for and against the causal hypothesis, no significant correlation was found between effect size and sample size. Nor did the Tandem Procedure find evidence for publication bias. Nonetheless, it does appear that the conclusions of the original meta-analysis

(Greitemeyer & Mügge, 2014) of consistent results may, in fact, be inconsistent with the pattern of effect sizes from individual studies.

Regarding the R-index, results indicated a general pattern of underpowered studies, most often with null results. Success rate was only 0.285 with observed power of 0.271 and R-index of 0.226. Such results suggest that this pool of data is not, overall, able to support links between violent video games and reduced prosocial behavior. Changing the criterion for statistical significance to 0.10 did not alter this scenario. Even eliminating smaller studies (<100) did not change this picture significantly either. In studies with $n > 100$, the correlation between sample size and effect size was $r = -0.27$, with a success rate of 0.375 with observed power of 0.223, and R-index of 0.290. R-index results in this range are generally consistent with null results but in which some studies may be reported significant results where null results may have gone unreported.

In the current study, since publication bias was not observed, and *p*-values were heterogeneous, p-curve analyses were not conducted.

3.3. Discussion

Overall, these results suggest that studies with null results are quite common in this field. Unlike the "best practices" studies of violent games and aggression which suggested a generally consistent but underpowered and difficult to replicate field, reported results in the prosocial realm appear to be more authentic, despite also being underpowered. It is important to point out that even some of the "statistically significant" results reported in the original meta-analysis have subsequently proven to be difficult to replicate (e.g. Tear & Nielsen, 2013, 2014). Thus, although selective reporting of results appears to be less an issue for this field than for the aggression field, caution is still warranted in regards to making causal attributions about video game influences based on this pool of data.

4. Study 3

Studies one and two concerned themselves with the issue of false positives in violent video game research. Given that many studies include relatively small samples, and that null results can be difficult to interpret, there is also a potential for false negatives to arise in the literature, namely studies that purport to find evidence against a relationship when the findings are, in fact, Type II error. Thus study 3 will concern itself with this issue. A pool of recent experimental studies with null results were examined using Bayes Factors to determine the degree to which such studies do, in fact, support the null.

4.1. Included studies

PsychINFO was searched using the terms "video games" AND "violence" AND "experiment" the first two as subject terms, the last as "all text." This resulted in 35 studies being identified. Abstracts of the studies were then examined for purported null results, narrowing down the pool to 18 studies of video game violence from 2007 on. Restricting the date to approximately 2007 was done for two reasons. First, this allowed the analysis to focus on the most recent up-to-date science, rather than older studies (many of which used "violent" video games no longer considered threatening to society such as *Zaxxon* or *Centipede*.) The period around 2007 also saw a relatively proliferation of studies with null results which have continued to the present time, providing an adequate and temporally homogenous pool of studies to consider. Authors of the identified studies were contacted for additional information needed to calculate Bayes Factors. Data from one study was no longer available, and several other authors did not respond to requests for data resulting in a final pool of 14 studies. These are listed in Appendix A.

4.2. Calculation of Bayes factors

Bayes Factors are numerical calculations that, rather than a binary decision as in traditional null-hypothesis significance testing, allow examiners to gauge relative support for null and alternative hypotheses. They are in effect, a ratio of the likelihood of the data given a given theory (H_a) and an alternate theory, typically the null hypothesis (H_o). This ratio allows for an examination of the data in relation to both a given theory and the null hypotheses, offering potential support for the null, which is difficult under traditional null hypothesis significance testing (NHST). Bayes factors are not a panacea for the null aversion issue influencing behavioral science. Bayes factors, like NHST, are sensitive to sample size, and may be "hacked" through QRPs as with NHST. Further, support for the null may be difficult in larger samples, or when the target value for H_a is set low (in effect, weaker theories are more difficult to disprove). And differing calculation methods may be more conservative with regards to support for the null (Dienes, 2015). Thus, Bayes factors are an improvement on traditional NHST but can't supplant careful analyses and conservative interpretation of effect sizes.

Bayes factor analyses in the current study were based on the approach of Dienes (2014). This approach is a particular conservative approach regarding support for H_o. Through such an approach the observed results from a given study can be compared against the results expected from a given theory. The predicted effect sizes are typically unknown and unclear, but the results from meta-analyses can be used when such meta-analyses are used by proponents of a theory as an index of "true" effects in the populations. Thus for the current study, meta-analytic results from Anderson et al. (2010) are used as the H_a benchmark.[1] With Bayes factor analyses the observed effect size, and the standard error of the study (which is sample size dependent) can be used to compute how likely the observed results were to be obtained given the assumption that either H_a or H_o are true. Represented as a ratio, figures greater than 1.0 provide increasing surety that the evidence support H_a whereas figures below 1.0 provide increasing evidence for H_o.

One study (Valadez & Ferguson, 2012) demonstrates some of the difficulty in interpreting effect sizes in this area. Unlike most experimental studies which use posttest only designs, Valadez and Ferguson pretested aggression prior to the experimental manipulation. Interestingly, aggression decreased across all video game conditions over time, whether violent or not. This raises the potential that experiments that do find a mean group difference may be incorrect in assuming that any such differences represent an increase in aggression. Instead, they may represent differential *declines* in aggression. While this may be interesting to know, it is very different in outcome from the hypothesis that violent games increase aggression. Thus, even positive BFs must be interpreted with care. However, this study was evaluated similarly to the other studies, considering posttest differences only.

4.3. Results

4.3.1. P-slacking

Mean sample sizes of studies with null results were examined in relation to sample sizes from meta-analyses asserting the presence of effects (e.g. Anderson et al., 2010; Greitemeyer & Mügge, 2014). The means sample size for studies with null results (M = 96.95) was between that for video games and aggression (Anderson et al., 2010; M = 93.07) and video games and prosocial behavior (Greitemeyer & Mügge, 2014; M = 107.43). The mean for the last meta-analysis was increased by a single outlier study ($n = 320$) without which the mean would have been 91.08. A one-way ANOVA on the sample sizes proved to be non-significant [$F(2, 57) = 0.22$, $p = 0.802$.] Thus there is little

[1] This should not be taken as an endorsement of this meta-analysis' results as accurate, only that they are used as population level benchmarks by advocates of the causal position.

evidence that studies with null results are deliberately underpowered to produce null effects.

4.3.2. Bayes factors

Results of the Bayes factor analyses are presented in Table 1. As can be seen, most studies that claim support for the null do, in fact, support the null. Two studies (Ivory & Kalyanaraman, 2007; Teng, Chong, Siew, & Skoric, 2011) reported initial inconsistencies in results (i.e. some outcomes in support of the alternative hypothesis, some outcomes in support of the null hypothesis), and results of the Bayes factor analyses supported these inconsistencies. Only a single study (Elson, Breuer, Van Looy, Kneer, & Quandt, 2015) initially reported as null evidenced some support for the alternative hypothesis through Bayesian analyses. However, this finding should be tempered by two observations. First, the Bayesian analysis is based on the use of the standardized version of the noise blast aggression measure used in the study (Ferguson et al., 2008). However, as the authors note, the field has not traditionally used the noise blast measure in a standardized fashion, with sometimes even the same lab changing the method for extracting aggression from one study to the next, a pattern often indicative of questionable researcher practices (see Ferguson, 2013 for discussion). As Elson and colleagues note (see also Elson, Mohseni, Breuer, Scharkow, & Quandt, 2014), using different approaches to extracting data from the noise blast measure from the same sample, it is possible to make it appear as if violent games increase aggression, decrease aggression or have no influence at all. Thus these findings for Elson et al. (2015) need to be tempered by the observation that they are not consistent across all ways in which the noise blast measure has been used in the literature. Second, a follow up study by the same research group produced results clearly supportive of the null hypothesis via Bayesian analyses (Kneer, Knapp, & Elson, 2014). Thus it should not be interpreted that this research group produces false negative results.

Table 1
Bayes factor analyses of studies with null results in violent video game research.

Study	Mean diff	Stnd error	BF	Outcome
Ferguson et al. (2008)	0.18	0.297	0.75	Null
Ferguson et al. (2008) (no choice)	−0.29	0.4044	0.37	Null
Ferguson & Rueda	−0.24	0.394	0.28	Null
Jerabeck & Ferguson	0.1	0.175	0.64	Null
Valadez & Ferguson	1.85	4.32	0.47	Null
Przybylski, Deci, Rigby, and Ryan (2014) study 1	0.009	0.201	0.26	Null
Przybylski et al. (2014) study 2	0.162	0.196	0.58	Null
Przybylski et al. (2014) study 5	0.047	0.192	0.3	Null
Tear and Nielsen (2013) study 1	0	0.11	0.17	Null
Tear and Nielsen (2013) study 2a	−0.13	0.168	0.18	Null
Tear and Nielsen (2013) study 2b	−0.25	0.151	0.06	Null
Tear and Nielsen (2013) study 3	−0.12	0.175	0.22	Null
Tear and Nielsen (2014) donation	−0.64	0.444	0.24	Null
Tear and Nielsen (2014) hurting	−0.059	0.356	0.43	Null
Tear and Nielsen (2014) helping	0.002	0.351	0.61	Null
Adachi and Willoughby (2011) study 1	0.17	0.5	0.88	Null
Adachi and Willoughby (2011) study 2	0.13	0.41	0.73	Null
Elson et al. intensity	0.35	0.193	4.55	Ha
Elson et al. duration	0.273	0.281	1.26	Ha
Ballard experimenter	−0.311	0.107	0.04	Null
Ballard partner	−0.536	0.119	0.03	Null
Eden Eschet	−5.35	1.94	0.04	Null
Ivory Hostility	0.227	0.161	1.16	Ha
Ivory Cognition	0.052	0.093	0.22	Null
Teng et al. aggression	−0.091	0.143	0.16	Null
Teng et al. hostility	−0.13	0.127	0.1	Null
Teng et al. ATVS	0.42	0.11	1036.59	Ha
Teng et al. empathy	−0.05	0.131	0.31	Null
Puri (2012)	−0.031	0.158	0.21	Null

Note: Ferguson (2008) (no choice) indicates analyses only for conditions where participants were randomized into a condition in which they were forced to play a particular game. Conditions in which participants were randomized to have the opportunity to choose a game were not included.

4.4. Discussion

Results from Bayes factor analyses generally supported the conclusion of existing null experimental studies of violent video game effects that find evidence for the null hypothesis. There is little evidence that null experimental studies are underpowered at least in relation to studies with statistically significant findings. And Bayes factors indicated general support for the null among this sample of studies. Thus, Type II error is not a sufficient explanation for observed null findings in the field of video game violence.

Thus results from this study confirm a consistent set of analyses which fail to confirm social cognitive theories linking violent video games to aggression. These studies are similar in size to those often used to highlight effects. Thus, it is clearly not possible to communicate that results from this field are uniform or robust.

5. General discussion

The issue of false positives and negatives in social science has gotten considerable attention in recent years. If published research results do not represent the full range of research studies actually conducted, knowledge transmission of a research field can become distorted. The current paper considered video game violence research as an example, with potential problems for this field likely illustrative of problems faced elsewhere, particularly fields that potentially overlap with more or political agendas, however well-meaning (e.g. spanking effects, racial or gender issues, etc.).

Three analyses examined the issue of whether previous studies of video game violence may have experienced either false positive or false negative results. Outcomes were mixed. The potential for false positives is most pronounced in studies that have been highlighted as linking violent video games to aggressive outcomes. Publication bias is clearly an issue for these studies, and they are producing positive results at a level higher than would be expected given their observed power. The issue for studies of violent video games and prosocial behavior are different. Little evidence of publication bias existed here, and results were widely mixed. However, results from these studies do not appear to be able to bear the burden of the causal conclusions made by some scholars. Lastly, little evidence emerged for the existence of a "false negative" problem among studies with null results. Null studies generally did, in fact, support the null, albeit some more strongly than others. Although video game experiments often draw from low-power samples, this was no more an issue for studies with null results than for other studies in this realm.

In general the summed research product of studies in this realm do not comport well with statements of consistent and important causal effects issued by some scholars (e.g. Anderson et al., 2010; Greitemeyer & Mügge, 2014). This may be due to the limited utility of meta-analysis to address inconsistencies in the research literature. In particular, mean effect sizes are likely of limited informational value given high study heterogeneity. To give an example, were 10 studies to test the hypothesis "X causes Y" with 5 of those studies returning mean effect sizes of $r = 0.20$, and the other 5 returning effect sizes of $r = 0.0$, it would be most valuable to try to understand why studies are returning inconsistent results. Merely meta-analyzing the effects to an average of $r = 0.10$ and asserting that this result is "true" on the population level, would be inappropriate.

At this juncture, that publication bias exists in studies of video game violence and aggression is reasonably clear. Publication bias was observed in study 1 despite the use of the Tandem Procedure, a very conservative approach to identifying publication bias. Indeed, underidentification of publication bias rather than overidentification is a greater likelihood for the Tandem Procedure given its conservative nature (Ferguson & Brannick, 2012). Publication bias was not indicated for studies of prosocial behavior, where many null studies exist. However, the problem for studies with null results appears to be that they are

not being effectively communicated to the scholarly community and general public.

At this juncture, the misuse of meta-analyses as "debate enders" or arbiters of population level "truth" should be regarded with extreme suspicion. The instability of meta-analyses are by now well known. Related issues include the sensitivity of meta-analyses to publication bias, with even relatively modest publication bias capable of producing spurious results (Scargle, 2000; Schonemann & Scargle, 2008). In many fields, including but not remotely limited to video games, competing scholars may release competing meta-analyses with differing conclusions, suggesting that meta-analyses likely fail at what is considered their primary function, namely adding objectivity to narrative reviews. However, given increasing reason to believe that between-study heterogeneity is more the norm than exception, this use of meta-analysis to obtain a purported population applicable effect size may simply be unwarranted.

The contention is not that video game violence research is uniquely problematic in regards to false positives, but rather that the issues seen for this field are likely common across many others, particularly but not limited to those with overlapping moral, political or social concerns. Understanding that the transmission of knowledge in social science and that flawed transmission of knowledge may produce informational biases can help us to correct these problems and create a more objective social science moving forward.

5.1. Suggestions for a road forward

One difficulty is that it is always unclear what to do with a field producing muddled results. How many null results are required before a theory is falsified? How do we treat null results…are they type II error or fatal to the theory in question? Do low power studies increase unreliability of both null and "statistically significant" findings?

With these questions in mind, several suggestions are offered that may help improve the interpretability and transparency of research results across social science. First, many scholars have recommended increased use of open science and pre-registration of studies. The movement toward open science remains in relative infancy in behavioral science, yet this may help us to put research results in a clearer framework and reduce the specter of QRPs which currently hangs over many areas of social science. Open science and preregistration has been advocated for social science generally (Carpenter, 2012), and results from preregistered trials have both supported and questioned some previously held ideas in behavioral science considered to be "true" (e.g. Klein et al., 2014; Lynott et al., 2014). Once pre-registered trials are developed, these can be compared to unregistered studies, to see if differences in effect are noticed. For instance, if pre-registered trials are more likely to produce null results than unregistered studies, this could be taken as evidence that positive results are due to QRPs and researcher expectancy effects. Several preregistered trials of video game violence effects have been produced (Ferguson et al., 2015; McCarthy, Coley, Wagner, Zengel, & Basham, 2016) with Bayesian analyses of null results, however more would certainly be welcome.

Second, and related, across fields social science would do well to increasingly adopt standardized outcome measures and measurement procedures. At present within aggression research, there are problems with research labs changing the way aggression is measured from one study to the next using the same instrument (see Elson et al., 2014 for discussion), or, in another case, operationalizing violent game exposure differently from one study to the next using the same dataset (e.g. Gentile et al., 2009, 2011; Gentile, Li, Khoo, Prot, & Anderson, 2014). This may be considered akin to chemists purposefully calibrating their instruments differently from one study to the next to achieve a desired outcome, then presenting the results as equivalent. However, it is unlikely that these issues are unique to video game research, and poor standardization is probably common aside from a few areas, such as clinical or I/O psychology where well developed and validated measures are commonly employed. Adopting clearly standardized instruments should be a priority for behavioral science. Further, the clinical utility and validity of measures needs to be determined before they are used to speak to clinically significant outcomes such as societally relevant aggression. Currently scholars are often *out on a limb* to the extent that they generalize poorly validated measures to serious aggression and violence. This may be true for other fields as well, insofar as scholars may be too quick to generalize results from esoteric laboratory measures to real-life.

Third, a movement away from traditional null hypothesis significance testing (NHST) toward alternative approaches would be desirable. Careful and conservative interpretation of effect sizes is currently lacking in the field, and too many "small is big" arguments are being used to exaggerate the impact of potentially trivial effects. The field needs to be able to come to a clearer identification of what effects are, ultimately, trivial. Or put clearer, agreed standards for theory falsification need to be clarified. Aside from effect sizes, use of Bayesian statistics may help to lend more clarity to null results. However, Bayesian statistics are potentially hackable in the same sense and manner as traditional NHST, such as through elimination of groups, altered analyses, eliminating covariates, etc.

Fourth, researchers need to increasingly include pretesting in research designs. Although testing and demand characteristic effects are a valid concern, without pretesting it is impossible to conclude that any differences at posttest are due to an increase (as in the example of aggression) as opposed to a differential decrease. Randomization should assure that pretest means across groups are about equal, but this does not inform whether those pretest means were higher or lower than the posttest means. Thus, if groups differ at post-test it is often *assumed* that one group *increased* in aggressiveness. However, it is plausible that both groups may have *decreased* in aggression, but with one group doing so more than the other. These are two very different outcomes that should not be interpreted as similar. Researchers have likely avoided pretesting due to concerns pretesting will conflict with the deception commonly used in aggression studies. As noted earlier, demand characteristics are certainly a valid concern. It may be possible to reduce this potential by including several distractor tasks between the pretest and experimental manipulation, then again between the experimental manipulation and posttest. Unfortunately few studies of video games use distractor tasks, which could be useful whether a pretest is employed or not. Again, this may not be an issue limited to video game research. Researchers may often infer mean differences indicate a change from pre to post, without actually having measured pre-scores. Without such pre-scores the direction of change cannot be inferred.

It would also be positive for researchers with varying views to find means to communicate and exchange ideas in open forums. Within video game research at present this has been accomplished only a single time (Ferguson & Konijn, 2015). In other fields as well, theoretical differences often lead to long-term acrimonious exchanges. With less personal investment at stake, and an atmosphere of collegial exchange, debates could become more informative and enlightening than acrimonious.

It is worth noting that the interplay between science and the general public, particularly on issues of particular social interest, is complex (Singh, Hallmayer, & Illes, 2007). This can be particularly true in a social/political environment in which concerns about a particular issue take on a "moral panic" tone as has been the case for video game violence (Bowman, 2016). Perceptions of truthiness in the public depend not only on the quality of data, but also which studies the media choose to report and how the public perceives them. Examples of this include the rise and fall of the lobotomy as a savior, then tormentor of the chronically mentally ill, as well as public skepticism of medical scientists' assurances of the safety of MMR vaccines, given previous false assurances on other issues such as mad cow disease. The interplay between science and news coverage of science can produce false

perceptions of scientific certainty in the mind of the public. This is particularly true on the issue of video games when "bad news" stories about video game effects get more coverage than do "good news" stories (Bowman, 2016).

5.2. Implications for law and public policy

The promulgation of false positive results can have tangible impacts on public policy both as endorsed by professional advocacy groups as well as for legal decisions. Although certainly not limited to the issue of video games, video game policy is once more instructive. For instance, in 2015 the American Psychological Association released a problematic and misleading resolution statement on video game violence, despite the appeal of over 230 scholars to avoid false positive public statements on the issue (Consortium of Scholars, 2013). The American Psychiatric Association does not have a policy statement on video games, although the American Academy of Child and Adolescent Psychiatry produces a series of "facts for families" pamphlets that, arguably, exaggerate the evidence for media effects, failing to inform families of studies with null results in this area. Such problematic public documents are arguably part of a larger culture of media effect exaggeration among professional advocacy organizations (Ferguson & Beresin, 2017).

The costs of false positive results can become evident in legal cases. Many legal cases may see the introduction of amicus briefs by professional advocacy organizations attempting to speak to the scientific accuracy of particular beliefs. If such amicus briefs, or other briefs citing professional advocacy organization resolution statements ultimately contain inaccurate false positive claims, these briefs could potentially mislead the court. Further, inaccurate public statements can also damage the reputation of the scientific field in the eyes of the court (Hall, Day, & Hall, 2011). This appears to have happened, for instance, in the case of Brown v EMA (2011). This US Supreme Court case considered the regulation of violent video game sales to minors and the research evidence cited to support such regulation. Opposing groups of scholars both supported and criticized the evidence base used to support regulation. The court in its majority opinion sided with the more skeptical view stating, "These studies have been rejected by every court to consider them, and with good reason…" and went on to echo the concerns about this body of research many scholars have also voiced.

It is clear that video game (and other media violence) research has entered a period where significant evidence of false positive results exists in some realms, and exaggerations of effects remain problematic (Markey, French, & Markey, 2015). Thus, it is recommended that policy makers considering legislation on the topic of media effects are unlikely to be able to count on a clear, consistent and high-quality evidence base on which to make policy decisions. Statements by some scholars and professional advocacy organizations may reflect political expediency and advocacy goals rather than objective overviews of the current state of scientific literature.

5.3. Conclusion

The field of video game violence continues to be limited by issues related to QRPs, difficulties in faithfully communicating inconsistent results, and ignorance of null effects and failed replications. A new commitment to open science, standardized methods, preregistration of studies and alternative models of analysis may help elucidate whether effects do or do not exist. Until more rigorous methods are adopted, debates related to cultural issues may continue to resemble culture war more than true scientific discussion.

The peccadillos of video game research are, in fact, probably not unique to the field. Indeed, although video game research is used here as an example of a larger problem, it's possible that fixing the problem in this field through a commitment to more rigorous methods, could also provide a road forward for other fields.

Appendix A

A.1. Reports included in Study 1

Anderson, C. A., & Carnagey, N. L. (2009). Causal effects of violent sports video games on aggression: Is it competitiveness or violent content?. *Journal Of Experimental Social Psychology*, *45*(4), 731–739. doi:https://doi.org/10.1016/j.jesp.2009.04.019

Anderson, C. A., & Dill, K. E. (2000). Video games and aggressive thoughts, feelings, and behavior in the laboratory and in life. *Journal of Personality and Social Psychology, 78*, 772–790.

Anderson, C. A., & Murphy, C.R. (2003). Violent video games and aggressive behavior in young women. *Aggressive Behavior, 29*, 423–429.

Anderson, C. A., Carnagey, N. L., Flanagan, M., Benjamin, A. J., Eubanks, J., & Valentine, J. C. (2004). Violent video games: Specific effects of violent content on aggressive thoughts and behavior. *Advances in Experimental Social Psychology, 36*, 199–249.

Anderson, C.A., Gentile, D.A., & Buckley, K.E. (2007). Violent Video Game Effects on Children and Adolescents: Theory, Research, and Public Policy. New York: Oxford University Press.

Arriaga, P., Esteves, F., Carneiro, P., & Monterio, M. B. (2008). Are the effects of Unreal violent video games pronounced when playing with a virtual reality system? *Aggressive Behavior, 34*, 521–538.

Ballard, M. E., & Lineberger, R. (1999). Video game violence and confederate gender: Effects on reward and punishment given by college males. *Sex Roles, 41*, 541–558.

Barlett, C. P., Branch, O., Rodeheffer, C., & Harris, R. J. (2009). How long do the short term video game effects last? *Aggressive Behavior, 35*, 225–236.

Bartholow, B. D., & Anderson, C. A. (2002). Effects of Violent Video Games on Aggressive Behavior: Potential Sex Differences. *Journal of Experimental Social Psychology, 38*, 283–290.

Bartholow, B. D., Sestir, M. A. & Davis, M. D. (2005). Correlates and consequences of exposure to videogame violence: Hostile personality, empathy, and aggressive behavior. *Personality and Social Psychology Bulletin, 31*, 1573–1586.

Brady, S.S., & Mathews, K.A. (2006). Effects of media violence on health-related outcomes among young men. *Archives of Pediatric & Adolescent Medicine, 160*, 341–347.

Carnagey, N. L., & Anderson, C.A. (2005). The effects of reward and punishment in violent video games on aggressive affect, cognition, and behavior. *Psychological Science, 16*, 882–889.

Cicchiriool, V. & Chory-Assad, R.M. (2006). Effects of affective orientation and video game play on aggressive thoughts and behaviors. *Journal of Broadcasting & Electronic Media, 49*, 435–449.

Gentile, D. A., Anderson, C. A., Yukawa, S., Ihori, N., Saleem, M., Ming, L. K., & … Sakamoto, A. (2009). The effects of prosocial video games on prosocial behaviors: International evidence from correlational, longitudinal, and experimental studies. *Personality And Social Psychology Bulletin*, *35*(6), 752–763. doi:https://doi.org/10.1177/0146167209333045

Irwin, A. R., & Gross, A. M. (1995). Cognitive tempo, violent video games, and aggressive behavior in young boys. *Journal of Family Violence, 10*, 337–350.

Katori, T. (2001). Bouryokuteki bideogemu no kougeki sokushin kouka to sougosayousei [The effects of violent video games and interactivity on aggression].Proceedings of the 42nd convention of the Japanese Society of Social Science, pp. 602–603.

Konijn, E. A., Bijnank, M. N., & Bushman, B. J. (2007). I wish I were a warrior: The role of wishful identification in the effects of violent video games on aggression in adolescent boys. *Developmental Psychology, 43*, 1038–1044.

Sakamoto, A, Kobayashi, S., & Mouri, M. (2001). Kougekigata terebigemu no shiyou ga joshi daigakusei no bouryokusei ni oyobosu eikyou: genjitsusei to houshousei no chousei kouka [The effect of violent video game use on violence of female university students: The adjustment effect of reality and reward].The Japanese Psychological Association 65th Annual Meeting, p. 804.

Sakamoto, A. Ozaki, M., Narushima, R., Mori, T., Sakamoto, K., Takahira, M., et al. (2001). Terebigemu asobi ga ningen no bouryoku koudou ni oyobosu eikyo to sono katei: Joshidaigakusei ni taisuru 2-tsu no shakaishinrigakuteki jikken [The influence of video game play on human violence and its process: Two social psychological experiments of female university students]. Studies in Simulation and Gaming, 11(1), 28–39.

Schutte, N. S., Malouff, J. M., Post-Gorden, J. C., & Rodasta, A. L. (1988). Effects of playing video games on children's aggressive and other behaviors. *Journal of Applied Social Psychology, 18*, 454–460.

Sheese, B. E. & Graziano W. G. (2005) Deciding to defect: The effects of video-game violence on cooperative behavior. Psychological Science 16, 354–357.

Yukawa, S., & Yoshida, F. (2000). Bouryokuteki terebigemu to kougekikoudou: gemu no seisitsu to insho oyobi sankasei no kouka [Violent video games and aggressive behavior: The effects of game format, impression and participation]. Proceeding of the 41st convention of the Japanese Society of Social Psychology, pp. 74–75.

A.2. Reports included in Study 2

Bösche, W. (2010). Violent video games prime both aggressive and positive cognitions. *Journal Of Media Psychology, 22*(4), 139–146. doi:https://doi.org/10.1027/1864-1105/a000019.

Bushman, B. J., & Anderson, C. A. (2009). Comfortably numb: Desensitizing effects of violent media on helping others. *Psychological Science, 20*, 273–277.

Gentile, D. A., Anderson, C. A., Yukawa, S., Ihori, N., Saleem, M., Ming, L., & … Sakamoto, A. (2009). The effects of prosocial video games on prosocial behaviors: International evidence from correlational, longitudinal, and experimental studies. *Personality And Social Psychology Bulletin, 35*(6), 752–763.

Gitter, S. A., Ewell, P. J., Guadagno, R. E., Stillman, T. F., & Baumeister, R. F. (2013). Virtually justifiable homicide: The effects of prosocial contexts on the link between violent video games, aggression, and prosocial and hostile cognition. *Aggressive Behavior*, *39*(5), 346–354. doi:https://doi.org/10.1002/ab.21487.

Greitemeyer, T. (2013). Playing video games cooperatively increases empathic concern. *Social Psychology*, *44*(6), 408–413. doi:https://doi.org/10.1027/1864-9335/a000154.

Greitemeyer, T., & Osswald, S. (2010). Effects of prosocial video games on prosocial behavior. *Journal Of Personality And Social Psychology, 98*(2), 211–221.

Greitemeyer, T., Osswald, S., & Brauer, M. (2010). Playing prosocial video games increases empathy and decreases schadenfreude. *Emotion, 10*(6), 796–802.

Greitemeyer, T., Traut-Mattausch, E., & Osswald, S. (2012). How to ameliorate negative effects of violent video games on cooperation: Play it cooperatively in a team. *Computers in Human Behavior, 28*(4), 1465–1470. doi: https://doi.org/10.1016/j.chb.2012.03.009.

Jerabeck, J. M., & Ferguson, C. J. (2013). The influence of solitary and cooperative violent video game play on aggressive and prosocial behavior. *Computers in Human Behavior, 26*, 2573–2578.

Saleem, M., Anderson, C. A., & Gentile, D. A. (2012). Effects of prosocial, neutral, and violent video games on college students' affect. *Aggressive Behavior, 38*(4), 263–271. doi:https://doi.org/10.1002/ab.21427.

Teng, S., Chong, G., Siew, A., & Skoric, M. M. (2011). Grand Theft Auto IV comes to Singapore: Effects of repeated exposure to violent video games on aggression. *Cyberpsychology, Behavior, And Social Networking, 14*(10), 597–602.

Whitaker, J. L., & Bushman, B. J. (2012). "Remain calm. Be kind." Effects of relaxing video games on aggressive and prosocial behavior. *Social Psychological And Personality Science, 3*(1), 88–92.

A.3. Reports included in Study 3

Adachi, P. 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*(4), 259–274. doi:https://doi.org/10.1037/a0024908.

Ballard, M., Visser, K., & Jocoy, K. (2012): Social context and video game play: Impact on cardiovascular and affective responses, *Mass Communication and Society, 15*:6, 875–898.

Eden, S., & Eshet-Alkalai, Y. (2014). The effect of digital games and game strategies on young adolescents' aggression. *Journal Of Educational Computing Research, 50*(4), 449–466. doi:https://doi.org/10.2190/EC.50.4.a.

Elson, M., Breuer, J., Van Looy, J., Kneer, J., & Quandt, T. (2015). Comparing apples and oranges? Evidence for pace of action as a confound in research on digital games and aggression. *Psychology Of Popular Media Culture*, *4*(2), 112–125. doi:https://doi.org/10.1037/ppm0000010.

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*(2), 99–108.

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

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*(3), 532–555. doi:https://doi.org/10.1111/j.1460-2466.2007.00356.x.

Jerabeck, J. M., & Ferguson, C. J. (2013). The influence of solitary and cooperative violent video game play on aggressive and prosocial behavior. *Computers in Human Behavior 29*(6), 2573–2578.

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*(3), 441–457. doi:https://doi.org/10.1037/a0034820.

Puri, K., & Pugliese, R. (2012). Sex, lies, and video games: Moral panics or uses and gratifications. *Bulletin of Science, Technology & Society, 32*, 345–352.

Tear, M. J., & Nielsen, M. (2014). Video games and prosocial behavior: A study of the effects of non-violent, violent and ultra-violent gameplay. *Computers In Human Behavior*, *41*8–13. doi:https://doi.org/10.1016/j.chb.2014.09.002.

Tear, M., & Nielson, M. (2013). Failure to demonstrate that playing violent video games diminishes prosocial behavior. *PLoS One*,*8*(7), e68382.

Teng, S. Z., Chong, G. M., Siew, A. C., & Skoric, M. M. (2011). Grand Theft Auto IV comes to Singapore: Effects of repeated exposure to violent video games on aggression. *Cyberpsychology, Behavior, And Social Networking*, *14*(10), 597–602. doi:https://doi.org/10.1089/cyber.2010.0115.

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.

References

Adachi, P. C., & Willoughby, T. (2011). The effect of violent video games on aggression: Is it more than just the violence? *Aggression And Violent Behavior*, *16*(1), 55–62.

Anderson, C. A., Shibuya, A., Ihori, N., Swing, E. L., Bushman, B. J., Sakamoto, A., ... Saleem, M. (2010). Violent video game effects on aggression, empathy, and prosocial behavior in eastern and western countries: A meta-analytic review. *Psychological Bulletin*, *136*(2), 151–173. https://doi.org/10.1037/a0018251.

Bargh, J., & Chartrand, T. (1999). The unbearable automaticity of being. *American Psychologist, 84*(7), 462–479.

Boot, W., Blakely, D., & Simons, D. (2011). Do action video games improve perception and cognition? *Frontiers in Psychology*, *2*, 226. https://doi.org/10.3389/fpsyg.2011.00226.

Bösche, W. (2010). Violent video games prime both aggressive and positive cognitions. *Journal of Media Psychology: Theories, Methods, and Applications*, *22*(4), 139–146. https://doi.org/10.1027/1864-1105/a000019.

Bowman, N. D. (2016). The rise (and refinement) of moral panic. In R. Kowert, & T. Quandt (Eds.), *The video game debate: Unraveling the physical, social, and psychological effects of digital games* (pp. 22–38). New York: Routledge.

Breuer, J., Vogelgesang, J., Quandt, T., & Festl, R. (2015). Violent video games and physical aggression: Evidence for a selection effect among adolescents. *Psychology Of Popular Media Culture*, *4*(4), 305–328. https://doi.org/10.1037/ppm0000035.

Brown v EMA (2011). Retrieved 7/1/11 from: http://www.supremecourt.gov/opinions/10pdf/08-1448.pdf.

Carpenter, S. (2012). Psychology's bold initiative. *Science*, *336*(6076), 1558–1561. https://doi.org/10.1126/science.335.6076.1558.

Charles, E., Baker, C., Hartman, K., Easton, B., & Kretzberger, C. (2013). Motion capture controls negate the violent video game effect. *Computers in Human Behavior, 29*, 2519–2523.

Consortium of Scholars (2013). Scholar's open statement to the APA task force on violent media. Retrieved from: http://www.scribd.com/doc/223284732/Scholar-s-Open-Letter-to-the-APA-Task-Force-On-Violent-Media-Opposing-APA-Policy-Statements-on-Violent-Media.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. https://doi.org/10.3389/fpsyg.2014.00781.

Dienes, Z. (2015). Representing H1 with flexible maxima, and a comparison of the Rouder et al (2009) and Dienes (2008) calculators. Retrieved from: http://www.lifesci.sussex.ac.uk/home/Zoltan_Dienes/inference/Bayes%20factor%20with%20t%20distribution.html.

Dominick, J. R. (1984). Videogames, television violence, and aggression in teenagers. *Journal of Communication, 34*, 136–147.

Doyen, S., Klein, O., Pichon, C. -L., & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS One*, *7*(1), e29081. https://doi.org/10.1371/journal.pone.0029081.

Elson, M., Breuer, J., Van Looy, J., Kneer, J., & Quandt, T. (2015). Comparing apples and oranges? Evidence for pace of action as a confound in research on digital games and aggression. *Psychology of Popular Media Culture*, *4*(2), 112–125. https://doi.org/10.1037/ppm0000010.

Elson, M., Mohseni, M., Breuer, 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*. https://doi.org/10.1037/a0035569.

Ferguson, C. J. (2013). Violent video games and the supreme court: Lessons for the scientific community in the wake of Brown v EMA. *American Psychologist*, *68*(2), 57–74.

Ferguson, C. J., & Beresin, E. (2017). Social science's curious war with pop culture and how it was lost: The media violence debate and the risks it holds for social science. *Preventive Medicine*, *99*, 69–76.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120–128.

Ferguson, C. J., & Konijn, E. A. (2015). She said/he said: A peaceful debate on video game violence. *Psychology of Popular Media Culture*, *4*(4), 397–411.

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

Ferguson, C. J., Trigani, B., Pilato, S., Miller, S., Foley, K., & Barr, H. (2015). Violent video games don't increase hostility in teens but they do stress girls out. *Psychiatric Quarterly*, *87*(1), 49–56.

Gentile, D., Li, D., Khoo, A., Prot, S., & Anderson, C. (2014). Mediators and moderators of long-term effects of violent video games on aggressive behavior practice, thinking, and action. *JAMA Pediatrics*. https://doi.org/10.1001/jamapediatrics.2014.63.

Gentile, D. A., Anderson, C. A., Yukawa, S., Saleem, M., Lim, K. M., Shibuya, A., ... Sakamoto, A. (2009). The effects of prosocial video games on prosocial behaviors: International evidence from correlational, longitudinal, and experimental studies. *Personality and Social Psychology Bulletin*, *35*, 752–763.

Gentile, D. A., Choo, H., Liau, A., Sim, T., Li, D., Fung, D., & Khoo, A. (2011). Pathological video game use among youths: A two-year longitudinal study. *Pediatrics*, *127*(2), e319–e329. https://doi.org/10.1542/peds.2010-1353.

Graybill, D., Kirsch, J., & Esselman, E. (1985). Effects of playing violent versus nonviolent video games on the aggressive ideation of aggressive and nonaggressive children. *Child Study Journal*, *15*(3), 199–205.

Greitemeyer, T., & Mügge, D. O. (2014). Video games do affect social outcomes: A meta-analytic review of the effects of violent and prosocial video game play. *Personality and Social Psychology Bulletin*, *40*(5), 578–589. https://doi.org/10.1177/0146167213520459.

Greitemeyer, T., Traut-Mattausch, E., & Osswald, S. (2012). How to ameliorate negative effects of violent video games on cooperation: Play it cooperatively in a team. *Computers in Human Behavior*, *28*(4), 1465–1470. https://doi.org/10.1016/j.chb.2012.03.009.

Griffiths, M. (2015). Video game bans: The debate about guns, GTA, and real-life violence. *The Independent* (Retrieved from: http://www.independent.co.uk/life-style/gadgets-and-tech/gaming/video-game-bans-the-debate-about-guns-gta-and-reallife-violence-10057296.html.

Hall, R., Day, T., & Hall, R. (2011). A plea for caution: Violent video games, the supreme court, and the role of science. *Mayo Clinic Proceedings*, *86*(4), 315–321.

Ioannidis, J. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*(6), 645–654. https://doi.org/10.1177/1745691612464056.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124.

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, *4*, 245–254.

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*(3), 532–555. https://doi.org/10.1111/j.1460-2466.2007.00356.x.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. J., Bahník, Š., Bernstein, M. J., ... Nosek, B. A. (2014). Investigating variation in replicability: A 'many labs' replication project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178.

Kneer, J., Knapp, F., & Elson, M. (2014). Challenged by rainbows: The effects of displayed violence, difficulty, and game-performance on arousal, cognition, aggressive behavior, and emotion. *Paper presented at the 64th annual conference of the international communication association, Seattle, WA*.

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0105825.

Lynott, D., Corker, K. S., Wortman, J., Connell, L., Donnellan, M. B., Lucas, R. E., & O'Brien, K. (2014). Replication of 'Experiencing physical warmth promotes interpersonal warmth' by Williams and Bargh (2008). *Social Psychology*, *45*(3), 216–222. https://doi.org/10.1027/1864-9335/a000187.

Markey, P. M., French, J. E., & Markey, C. N. (2015). Violent movies and severe acts of violence: Sensationalism versus science. *Human Communication Research*, *41*(2), 155–173. https://doi.org/10.1111/hcre.12046.

McCarthy, R. J., Coley, S. L., Wagner, M. F., Zengel, B., & Basham, A. (2016). Does playing video games with violent content temporarily increase aggressive inclinations? A pre-registered experimental study. *Journal of Experimental Social Psychology*, 6713–6719. https://doi.org/10.1016/j.jesp.2015.10.009.

Pashler, H., Coburn, N., & Harris, C. R. (2012). Priming of social distance? Failure to replicate effects on social and food judgments. *PLoS One*, *7*(8), e42510. https://doi.org/10.1371/journal.pone.0042510.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536. https://doi.org/10.1177/1745691612463401.

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*(3), 441–457. https://doi.org/10.1037/a0034820.

Scargle, J. D. (2000). Publication bias: The "file-drawer" problem in scientific inference. *Journal of Scientific Exploration*, *14*, 91–106 (Retrieved from http://www.scientificexploration.org/journal/jse_14_1_scargle.pdf).

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*(4), 551–566. https://doi.org/10.1037/a0029487.

Schimmack, U. (2014). Quantifying replicability: The R-index. Retrieved from: http://www.r-index.org/.

Schonemann, P. H., & Scargle, J. D. (2008). A generalized publication bias model. *Chinese Journal of Psychology*, *50*, 21–29 (Retrieved from http://www.schonemann.de/pdf/91_Schonemann_Scargle.pdf).

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, *9*(6), 666–681. https://doi.org/10.1177/1745691614553988.

Singh, J., Hallmayer, J., & Illes, J. (2007). Interacting and paradoxical forces in neuroscience and society. *Nature Reviews Neuroscience*, *8*, 153–160.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside bias, rational thinking, and intelligence. *Current Directions in Psychological Science*, *22*(4), 259–264. https://doi.org/10.1177/0963721413480174.

Tear, M., & Nielsen, M. (2013). Failure to demonstrate that playing violent video games diminishes prosocial behavior. *PLoS One*, *8*(7), e68382.

Tear, M. J., & Nielsen, M. (2014). Video games and prosocial behavior: A study of the effects of non-violent, violent and ultra-violent gameplay. *Computers in Human Behavior*, 8–13. https://doi.org/10.1016/j.chb.2014.09.002.

Teng, S. Z., Chong, G. M., Siew, A. C., & Skoric, M. M. (2011). Grand theft auto IV comes to Singapore: Effects of repeated exposure to violent video games on aggression. *Cyberpsychology, Behavior, and Social Networking*, *14*(10), 597–602. https://doi.org/10.1089/cyber.2010.0115.

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.

Vieira, E. (2014). The relationships among girls' prosocial video gaming, perspective-taking, sympathy, and thoughts about violence. *Communication Research*, *41*(7), 892–912.