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