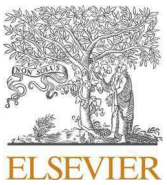
Journal of Experimental Social Psychology 93 (2021) 104082



Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Maximal positive controls: A method for estimating the largest plausible effect size

Joseph Hilgard

Illinois State University, DeGarmo Hall Room 435, Campus Box 4620, Normal, IL 61790-4620, United States of America

ARTICLE INFO

Keywords:
Violent video games
Aggression
Aggressive thought
Positive controls
Scientific self-correction

ABSTRACT

Effect sizes in social psychology are generally not large and are limited by error variance in manipulation and measurement. Effect sizes exceeding these limits are implausible and should be viewed with skepticism. Maximal positive controls, experimental conditions that should show an obvious and predictable effect, can provide estimates of the upper limits of plausible effect sizes on a measure. In this work, maximal positive controls are conducted for three measures of aggressive cognition, and the effect sizes obtained are compared to studies found through systematic review. Questions are raised regarding the plausibility of certain reports with effect sizes comparable to, or in excess of, the effect sizes found in maximal positive controls. Maximal positive controls may provide a means to identify implausible study results at lower cost than direct replication.

In the inferential statistics used in most psychology research, any effect size that is significantly different from zero is considered a success. Generally, the bigger the effect size, the better: Compared to small effect sizes, large effect sizes yield smaller *p* values, explain more of the variance in outcomes, indicate that smaller sample sizes are sufficient in power analysis, and suggest better cost-benefit ratios in interventions.

There are circumstances, however, in which an effect size is *too* large. Some effect sizes are so big that they could not be plausibly created by the hypothesized process that the study is intended to measure. Such an effect size fits neither the null nor the alternative hypothesis; more plausible is some third cause, such as an error in research design, data quality, or analysis. It is even possible that an unusually large effect size is caused by fabricated data. When researchers are asked to fabricate data, they tend to generate data showing much larger effect sizes than those found in genuine data (Akhtar-Danesh & Dehghan-Kooshkghazi, 2003; Hartgerink et al., 2019). The fraud of Diederik Stapel was discovered, in part, by the massive effect sizes reported in his studies (Levelt, 2012). Other studies have been scrutinized and retracted on the basis of implausibly large effect sizes and variance explained (e.g., Sosso et al., 2020; see also the work of Hans Eysenck and Ronald Grossarth-Maticek as criticized by Pelosi & Appleby, 1992).

One challenge in criticizing a study's reported effect as being "unusually large" is that such judgments can be subjective. Where a skeptic might judge an effect as implausibly large, doubting the quality and fidelity of the data, a proponent might judge the same effect as plausible, providing welcome support to a favored hypothesis or theory. How can these differences of opinion be resolved? In the absence of data, this is difficult. However, an estimate of the largest plausible effect size on the measure, an effect that logically should be strictly larger than the study's effect size, could provide a benchmark against which to compare study effects. It could be agreed that effects in excess of this largest plausible effect size are in error. In this work, I introduce the use of *maximal positive controls* as a way to establish the largest plausible effect size on a measure.

1. Introduction

1.1. Positive controls

Most researchers are familiar with *negative controls*, conditions in which the active component of the manipulation is omitted. For example, if a researcher tests the effects of violent video games on aggressive behavior, their treatment condition might have participants play a violent video game, and the negative control condition might have participants play a nonviolent video game. The use of a negative control allows estimation of the effect of the intervention (e.g., violent game content).

Less commonly used is the *positive control*, a condition in which participants receive a treatment that should have an obvious and predictable effect on the outcome. Positive controls allow researchers to test whether their methods are appropriately sensitive (Moery & Calin-Jageman, 2016). If the experiment is conducted, and an effect of the

E-mail addresses: jhilgard@gmail.com, jbhilga@ilstu.edu.

https://doi.org/10.1016/j.jesp.2020.104082
Received 23 August 2019; Received in revised form 2 November 2020; Accepted 3 November 2020
Available online 24 November 2020
0022-1031/

positive control is not observed, it suggests that there may be some flaw in the methods or measures. For example, a biologist might include in her assay a sample with a known quantity of the element she hopes to detect. If the other samples come up blank, but the positive control provides the correct reading, she can rule out equipment failure as an alternative explanation; the experimental samples are truly devoid of the expected element. On the other hand, if the element is not detected in the positive control sample, then she knows something is wrong with her measurements. Similarly, in a psychological experiment, a participant might first use a question to rate some obvious examples before rating the more nuanced stimuli of research interest. For example, a participant might first rate the naturalness of "a plastic toy model of a car" and "a tree growing on a mountain peak that has never been visited by humans" before rating the naturalness of spring water with added minerals (Rozin, 2006). If these positive control questions fail to show the anticipated large effect, it suggests a problem: perhaps participants are misunderstanding the question or withholding effort, or perhaps there has been some error in data analysis.

Although the use of positive controls in psychology is still in its infancy, one novel application is the use of positive controls in replication studies as a way to check the quality of the data collection. For example, when Moery and Calin-Jageman (2016) attempted to replicate the effects of exposure to organic food on prosocial behavior, they collected a test of the retrospective gambler's fallacy, an effect found to be replicable in previous large-sample, multi-site work. Although they could not replicate the effect of organic food stimuli on prosocial behavior, they did replicate the retrospective gambler's fallacy. The success of the positive control suggested that the failure to replicate the effect of organic food on prosocial behavior was not due to some broader issue with participant inattention or errors in data analysis.

Positive controls can be selected to anticipate a small, medium, or large effect size. Here I propose *maximal positive controls*, a class of positive control designed to find the largest plausible effect sizes on a given measure by minimizing sources of error variance. When a maximal positive control is easier to collect than a full-fledged replication, the collection of a maximal positive control can provide a cost-effective way to inspect the plausibility of a result.

1.2. Maximal positive controls

An experiment is typically designed to test whether some manipulation influences some outcome. If there is such an effect of the manipulation, rarely does it account for all the variance in the outcome, as even the strongest relationship is often attenuated by measurement error. Indeed, not every experiment yields significant results: Often there is no association, or the association is undetectably small. A maximal positive control makes the association as big as possible. Reported effect sizes exceeding this "big as possible" threshold may be in error.

Let us consider two ways to think about an experiment's effect size. First, we might think of the effect size as describing the ratio of variance between groups (variance explained) to variance within groups (error variance). In this context, an exceedingly large effect size indicates an unusually large difference between groups or an unusually small amount of variance within groups.

A second way to think about an experiment's effect size is to think of an experiment as being like a path model. A participant receives (or does not receive) some manipulation, which should influence some internal state of the participant, which is hypothesized to influence the participant's behavior or self-reports. Viewed as a path model, each process in this causal chain is one path between variables: from the manipulation to the internal state, and from the internal state to the dependent variable. Because the magnitude of each path is almost always less than a perfect 1.0, each step in the causal chain will have less and less explained variance and more and more error variance.

To increase variance between groups, a researcher could use stronger manipulations, test more obvious processes, or remove some steps from the causal chain. Suppose we are conducting an experiment testing how the persuasiveness of an essay is influenced by the author's credentials. Participants read an essay arguing in favor of some policy, and the essay is labeled as coming from either the social science department of a prestigious university (strong credential condition) or a layperson's Facebook page (weak credential condition). We could increase the effect size in several ways. First, we could increase the strength of the credential manipulation by labeling the essay with still more (and still less) credible sources. Second, we could confound together multiple factors that should influence the essay's persuasiveness: In addition to manipulating the credentials, we could also manipulate its quality of presentation and logical coherence of arguments, reasoning that the combined effect of all three factors should be strictly greater than the effect of credentials alone. Third, we could remove steps from the causal chain or replace weaker causal processes with stronger causal processes. For example, we could ask participants to indicate the perceived strength of the author's credentials, removing the causal steps in which the author's argument influences (or fails to influence) the participant's opinion.

To reduce variance within groups, one might ask all participants to make the same response. While this might not be particularly enlightening for questionnaire data, it can be interesting when examining behavior in some task. For example, a research report described an experiment in which participants played a video game for five minutes as a heroic or villainous avatar, then aggressed against another person by pouring them an amount of hot sauce to eat (Yoon & Vargas, 2014). The variance within condition seemed remarkably small for a process as potentially messy and unpredictable as pouring hot sauce without the aid of a pipette. A follow-up study explicitly instructed participants to all pour the same amount of hot sauce (Hilgard, 2019). The standard deviation of these deliberately-similar pours was greater than the reported standard deviation of pours from participants given no such instructions. This suggested that there was some error in the original research.

In studying the literature on aggressive cognition, I have seen some reported results that are of surprising magnitude. Let us apply the approach of maximal positive controls to three measures of aggressive cognition: the story completion task, the word completion task, and the aggressive-emotion Stroop.

2. Study 1: The story completion task

Aggressive thoughts are expected to cause aggressive behaviors. Studies of possible aggression-inducing stimuli such as violent media often use measurements of aggressive thought as outcomes. It is assumed that increases in measured aggressive thought will predict increases in aggressive behavior, but this predicted mediation is rarely demonstrated in the published literature.

One such rare demonstration of this mediation is provided by Hasan, Bègue, Scharkow, and Bushman (2013). In this study, participants played a violent or nonviolent game for 20 min a day on three consecutive days. After each gameplay session, their aggressive thoughts and behaviors were measured. Aggressive thoughts were measured using the story completion task, which asks participants to read a story stem and indicate ways in which the story might continue (C. A. Anderson, 1999a). Participants who expect the stories' characters to behave aggressively are said to have a "hostile expectation bias," an expectation that real-world interactions will end in aggression. Aggressive behavior was measured using the Taylor aggression paradigm, in which participants choose the intensity and duration of a loud, unpleasant noise to be sent to an opponent.

Hasan and colleagues reported a significant effect of violent video games on aggressive behavior, mediated, as predicted, by aggressive thoughts. Moreover, those effects accumulated across testing sessions. By the third day, participants who played a violent game behaved more aggressively ($d = 1.52$ [0.99, 2.05]) and thought more aggressively ($d =$

3.46 [2.72, 4.21]) than those who played nonviolent racing games. By the third day, random assignment to game explained 77.60% of variance in aggressive thoughts—a remarkable accomplishment for a manipulation lasting 1.4% of the day. Hasan and colleagues report similarly remarkable effects of violent video games on story completion task scores in another article (Hasan, Bègue, & Bushman, 2012, $d = 1.92$) and in a conference abstract (Hasan, Bègue, & Bushman, 2015, $d = 2.98$).

There is reason to believe these effects are too large to fit the alternative hypothesis. Effects of this size tend to be so large as to be obvious to casual observation. For example, you may have noticed that men tend to be taller than women; this is a large effect of $d = 1.85$ (Simmons, Nelson, & Simonsohn, 2013). If one hour of violent games divided across three days caused such dramatic changes in aggressive thoughts and behavior, we would notice whenever our friends or students purchased a new violent video game.

Effects in social psychology are typically much more subtle, averaging $d = 0.4$ and rarely exceeding $d = 1$. On average, they explain 10% or less of variance (Richard, Bond, & Stokes-Zoota, 2003). In meta-analysis, the effects of video games on similar measures of aggressive cognition explain about 6% of the variance, on average (Anderson et al., 2010; Greitemeyer & Mügge, 2014). It is therefore highly unusual that this one factor should explain 78% of the variance, particularly when one might expect scores to be influenced by measurement error, personality, or unique experiences taking place between sessions outside the study. Additionally, there should be variance due to stimulus: the story completion task consists of three different scenarios, counterbalanced across participants. If some scenarios elicit more aggressive responses than others, averaging across scenarios as Hasan et al. (2013) did will fold these effects of scenario into the error variance.

The article has had noticeable influence in aggression research. It has been cited more than a hundred and fifty times, has been included in three meta-analyses (Bushman, 2016; Calvert et al., 2017; Greitemeyer & Mügge, 2014), and was cited in the current APA policy statement on violent video games (American Psychological Association, 2015). Given this influence, a double-check could be valuable.

To conduct a direct replication of this work would have been prohibitively expensive, given that participants would need to return to the laboratory for three consecutive days. Instead, I checked the plausibility of these results by using a maximal positive control to estimate the largest plausible effect size on the story completion task. Rather than closely replicating the methods of the original study to try to estimate the precise causal effect of video game violence on story completion task scores, this approach instead uses deliberately dissimilar methods to estimate the largest plausible effect size on the story completion task.

In the original study, participants were asked to indicate how *a normal person* might behave in the story completion task scenarios. The causal process is presumed to be roughly as follows: Participants play the video game, experiencing the game's content. The game's content influences the participants' perceptions of how people generally behave (e.g., by activating aggressive schemas or teaching a hostile expectation bias). The participant's hostile expectation bias influences their responses on the story completion task.

In the present study, participants were asked to indicate how *the video game character* might behave in the task's scenarios. This simplified the causal process: Participants watched a video, experiencing the video's content. They then reported their experience of the video's content on the story completion task. This removed the step in which the game's content was expected to subtly influence the participant's social schema—an effect which, if it exists, should be of moderate size. Where Hasan et al. (2013) studied a subtle adjustment of social schemas through media influence, I studied a straightforward demand. This obvious manipulation indicated what might be, roughly, the largest effect one could expect on this measure.

2.1. Method

Planning for precision suggested a target sample size of approximately 50 per condition for a desired confidence interval width of 0.5 units of *d*. Participants were 142 college undergraduates (29 males, 113 females) participating for partial course credit. Simulations conducted after the experiment found that this sample size had 80% one-tailed power to detect a difference between that reported in Hasan et al. (2013) and an effect size in maximal positive control as large as $\delta = 2.75$; see the supporting materials for code. Most participants were White (66%), with an additional 22% identifying as Black, 1% as Asian, and 10% as another ethnicity. 24 participants identified as Hispanic. Participants came into the lab and were seated at a Qualtrics survey. Participants were warned that that they might view an objectionable or disturbing video; all provided informed consent (Illinois State University ethics board approval number IRB-2018-1144025).

Participants were randomly assigned to view one of three videos. In the violent condition, participants watched gameplay footage from *Grand Theft Auto V* (*GTAV*). In this video, the character Michael enters a strip club, shoots all the patrons and dancers to death, kills the cashier while she begs for her life, then goes outside into the street and engages the police in a prolonged shootout. I used this stimulus instead of the violent games from Hasan et al. (2013) (*Condemned 2, Call of Duty 4, The Club*) because the character is obviously antisocial rather than following military orders or acting in self-defense. In the racing control condition, participants watched gameplay footage of *DiRT 2*, one of the nonviolent games used in the control condition of Hasan et al. (2013). In this video, a man drives a 4-wheeler around a race track and performs stunts. In the peaceful control condition, participants watched gameplay footage from *Heavy Rain*. In this video, the character Ethan works at his drafting table, drawing an architectural sketch for a house. I added this condition because I was curious if this would yield even lower story completion scores than the racing control. Participants were instructed that they would "write what that video game character might do in a social situation" and that "There are no right or wrong answers. Just try to get a sense of the character and how they might act."

Following the video, participants performed the story completion task. Participants read three scenarios featuring the character from the video. In one, the character is rear-ended after stopping at a traffic light. In another, the character tries to persuade a friend to join him on a beach vacation. In a third, the character receives very poor service at a restaurant. Participants were asked to supply 20 completions total across three categories: things the character might do or say, things the character might think, or things the character might feel. These story stems are typical of the measure and are the same ones used in Hasan et al. (2013). Qualtrics automatically piped in the name of the character from the video.

Five research assistants coded the responses using the materials provided by C. A. Anderson (1999a). Instead of a dichotomous "not aggressive/aggressive" rule as in the original materials, responses were scored on a 1–7 scale from "not at all aggressive" to "highly aggressive" as performed by Hasan et al. (2013) (Hasan, personal correspondence, Oct 12, 2017). Coders rated each response on its own, blind to the participants' condition or other responses. Coders were discouraged from consulting each other about ambiguous cases, which can artificially increase intercoder reliability. Each provided response was scored as the average of coders' ratings. Scores were then averaged within each category (do or say, think, feel), then averaged within scenario, then averaged within subject.

To prepare the research assistants for coding, I held a training session in which we coded 27 randomly-selected responses together and discussed any differences of coding. Some research assistants were initially inclined to rate negative, non-aggressive feelings like "anxious" or "bad" as aggressive, but training corrected this tendency.

All data, materials, and code for this study and the following studies are available at https://osf.io/7um6d. I report all measures,

manipulations and exclusions. Sample size was determined in advance of any data analysis.

2.2. Analysis

I calculated the effect size as pairwise standardized mean differences between each condition. I did the same for the Day 3 data from Hasan et al. (2013) using raw data provided by the first author. I then compared the two effect sizes.

2.2.1. Intercoder reliability

Coders' scores showed excellent relative average consistency at the level of individual responses (ICC3k = 0.97). This consistency suggests minimal introduction of error and little attenuation of the effect size due to intercoder error.

2.2.2. Effects of scenario

A multilevel model with condition and scenario as fixed effects and participant as random effect was fit using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015). This found significant effects of condition ($\chi^2(2) = 71.40$, $p < .001$), scenario ($\chi^2(2) = 6.63$, $p = .036$), and a Condition × Scenario interaction ($\chi^2(2) = 62.00$, $p < .001$). Participants provided the least aggressive completions in the peaceful control condition, more aggressive completions in the racing control condition, and the most aggressive completions in the violent condition. Participants also provided the least aggressive completions in the beach vacation scenario, more aggressive completions in the restaurant scenario, and the most aggressive completions in the car accident scenario. These effects interacted such that the differences between scenarios were larger in conditions with more aggressive characters.

2.2.3. Comparisons to data from Hasan et al. (2013)

Means and standard deviations in the present data and in that of Hasan et al. (2013) are presented in Table 1. Individual participant scores are plotted in Fig. 1. The data from Hasan et al. show substantial separation of means and small standard deviations around those means, an effect of $d = 3.46$ [2.72, 4.21]. In the present data, the means were all much closer to 1, even for the violent-character condition. The SDs were also smaller, perhaps due to floor effect.

2.2.4. Comparing the effect sizes

Contrasting the violent condition against the racing control condition in the present data yielded an effect size $d = 1.71$, 95% CI [1.24, 2.18]. By contrast, the effect size reported in Hasan et al. (2013), day 3, was more than twice as large, $d = 3.46$ [2.72, 4.21].

The statistical significance of this difference can be tested with a *t*-test. The numerator is the difference between effect size estimates, and the denominator is a function of the standard errors and sample sizes of each effect size estimate. Because the two studies have different sample sizes, and thus, different standard errors, I used Welch's *t*-test, which is more reliable when the two samples have different standard errors. A one-tailed test was used given my hypothesis that the effect size would be larger in Hasan et al. (2013) than in the present data. The difference was statistically significant, $t(108.67) = 5.66$, $p < .001$.

Returning to the present data, the racing control condition provided more aggressive story completions than the peaceful control condition, $d = 1.32$, 95% CI [0.87, 1.76]. Contrasting the violent condition against the peaceful control condition yielded a larger effect size of $d = 2.51$, 95% CI [1.97, 3.04]. Nevertheless, a one-tailed Welch's *t*-test found that the effect reported by Hasan et al. (2013) was larger still, $t(118.52) = 2.94$, $p = .002$.

Across the three conditions in the current study, random assignment to condition explained $R^2 = 59.30\%$ of the variance in story completion task scores. In the data from Hasan et al. (2013), random assignment to condition explained $R^2 = 77.60\%$ of variance in story completion task scores.

Table 1
Descriptive statistics of the current data and that of Hasan et al. (2013).

Dataset	Condition	M	SD	N
My data	PeacefulControl	1.35	0.19	48
My data	RacingControl	1.67	0.28	47
My data	Violent	2.48	0.61	47
Hasan data	RacingControl	2.54	1.12	35
Hasan data	Violent	5.84	0.81	35

2.3. Other concerns about Hasan et al. (2013)

It is typical to test for hypothesis-awareness and failures of deception in studies such as these. Researchers are concerned that participants may realize that the game is expected to influence their behavior, or that they are not actually sending unpleasant stimuli to a real person, or that their behavior is being monitored and judged, all of which may lead to attempts to thwart the hypothesis.

I have found it to be difficult to routinely deceive participants; in a single-session experiment, up to 25% of participants might not be deceived (Hilgard, Engelhardt, Rouder, Segert, & Bartholow, 2019). However, Hasan and colleagues did not report a single savvy participant despite repeated presentation of a violent video game followed by the story completion task and the Taylor aggression paradigm. Under these conditions, it seems probable that at least one subject might suspect that the study is about the effects of violent video games on aggression and not "the effects of brightness of video games on visual perception" (Hasan et al., 2013, p. 225).

2.4. Systematic review

As an additional check, it is useful to consider what effect sizes are typically observed on the story completion task. I performed a systematic review and meta-analysis of experiments using the story completion task. I searched PsycINFO, PsycARTICLES, Web of Science, and Scopus using the terms *aggress** AND *story stem* and *aggress** AND *story completion*. 87 records were discarded as duplicates. 94 were excluded as irrelevant: typical reasons for exclusion at this stage included using a different measure (often the MacArthur Story Stems) or research design. One did not provide sufficient detail to calculate an effect size. An additional four studies were found through reference. Ten studies were retained for inclusion and coding, yielding 22 effect sizes. Effect sizes were calculated in terms of Hedges' *g* using reported means and SDs, test statistics, or a reported Pearson *r*. The sign of effects was set such that a hypothesis-consistent effect was given a positive sign (e.g., an increase given an aggressive stimulus or a decrease given a prosocial stimulus were both coded as positive effects). The distribution of effect size estimates is displayed in Fig. 2.

The average effect size was $g = 0.82$, 95% CI: [0.48, 1.16], a large effect. There was considerable heterogeneity, $I^2 = 93.60$, $\tau = 0.77$. In the above experiment, the observed maximal difference between an aggressive stimulus and a neutral control was $g = 1.69$. Of the three effect sizes exceeding this estimate, all were reported by Hasan and colleagues (Hasan et al., 2012, 2015, 2013). Studies in which Dr. Hasan served as the first author found significantly larger effects than did other studies, $z = 3.78$, $b = 1.16$, $p < .001$. On average, studies in which Dr. Hasan was not the first author found an effect that was 20.72% as large as the largest effect size in the present maximal positive control. Studies in which Dr. Hasan was the first author, on the other hand, found effects that were, on average, 71.48% as large as the largest effect size in the present maximal positive control. Excluding studies with Dr. Hasan as a first author reduced the average effect to $g = 0.48$, 95% CI: [0.34, 0.61], $I^2 = 50.16$, $\tau = 0.18$.

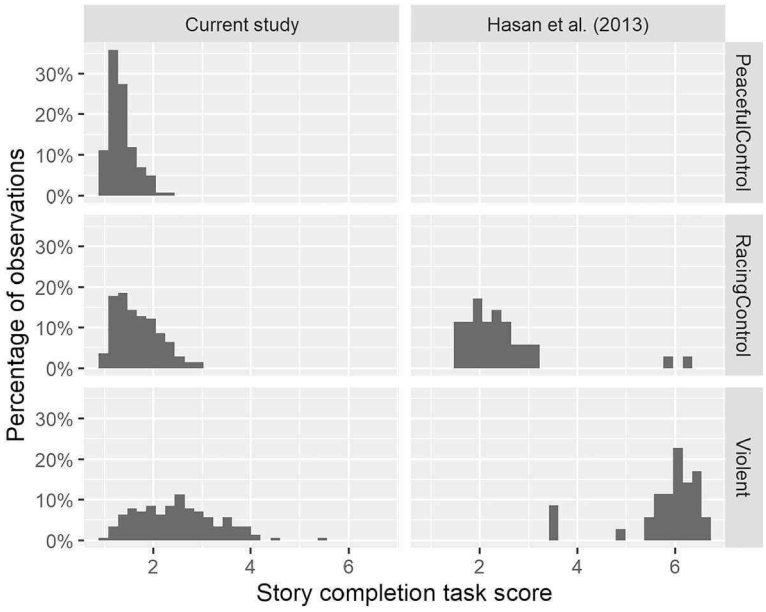


Fig. 1. Story completion task scores per participant in the present data and in Day 3 of Hasan et al. (2013). Data from Hasan et al. (2013) reveals higher mean scores and less overlap between distributions despite using a subtler manipulation. The peaceful control condition is new to this study and was not used in Hasan et al. (2013). Note that participant scores in my data represent an average across all three story scenarios, whereas those in Hasan et al. (2013) represent a single random scenario.

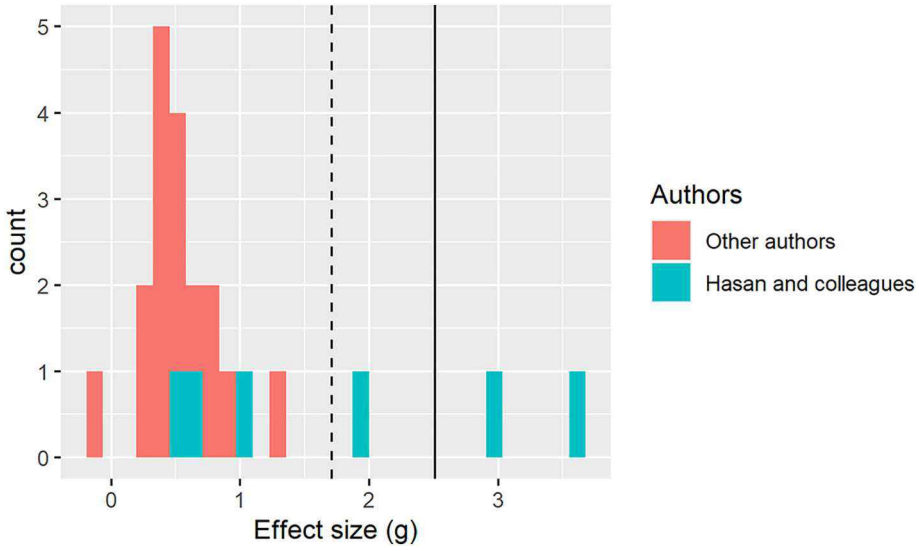


Fig. 2. Observed effect sizes in studies using the story completion task. The dashed vertical line indicates the effect size of the contrast between the mass shooter character from *GTAV* and the racer from *DiRT 2* in the present maximal positive control. The solid vertical line indicates the effect size of the contrast between the *GTAV* character and the peaceful architect from *Heavy Rain*. All estimates exceeding the effect sizes found in the present maximal positive control are reported by Hasan and colleagues.

2.5. Limitations

The guidelines for coding participant responses were a little vague. It is more common in the literature to score responses dichotomously as "aggressive" or "not aggressive" (e.g., Bushman & Anderson, 2002) using the scoring rubric provided by C. A. Anderson (1999a). When I asked how to score responses per Dr. Hasan's methods, Dr. Hasan said to code responses using the instructions from C. A. Anderson (1999a) but on a 1–7 scale instead of a 0, 1 scale. In the absence of a more rigorous coding rubric, it is hard to tell what distinguishes, for example, a rating of 3 from a rating of 4. Maybe differences in coding practices can explain some of the differences between Dr. Hasan's data and my own. I note that the modal code in Dr. Hasan's data is 2, whereas the modal code in my data is 1.

My approach to coding might have increased the variability across responses within a subject. It is possible that when all of a participants' responses are coded together some manner of Gestalt emerges and makes the ratings more like each other. Our approach was to avoid that Gestalt and code each response in isolation. If this increases variability within subjects, it will reduce the effect size.

Another limitation is the differences in participant demographics between the present research and Hasan et al. (2013). In Hasan et al. (2013), half of participants were male; in the present research, only 20% of participants were male. Additionally, research participants in Hasan et al. (2013) were French, whereas those in the present research were American. These differences in sample, however, are unlikely to explain the differences in results, as there is no theoretical reason to expect an interaction between condition and participant sex or nationality on story completion task scores of sufficient strength to explain the differences between the present study and Hasan et al. (2013).

Although these results are informative regarding the plausibility of the effect sizes reported in Hasan et al. (2013), it would be interesting to know whether maximal positive controls can be implemented in other similar measures to detect implausibly large effect sizes. It might also be interesting to explore the typical ratio between effect sizes in primary research and the effect size in maximal positive control. I continue this work in Study 2 by conducting a maximal positive control using another measure of aggressive cognitions, the word completion task (C. A.

Anderson, 1999b).

3. Study 2: The word completion task

The word completion task is another popular measure of aggressive cognition. Participants are presented with word stems that can be completed as either aggressive or non-aggressive words (e.g. KI_ could be completed as KILL or KISS; MU_ER could become MURDER or MUTTER). To the extent that a person's aggressive schemas are activated, it is expected that that person will generate more aggressive, and fewer non-aggressive, word completions. The task is sometimes scored as the total count of aggressive completions. At other times, the task is scored as the proportion of aggressive completions out of total completions.

Several extraneous factors may introduce error variance into the word completion task. Verbal skill may influence whether participants are able to think of a completion, aggressive or otherwise, for some word stems. This influence may be greater still when the task uses a time limit to encourage automatic responding. There may also be error variance in whether a given word-stem completion necessarily represents aggressive cognition. Some words like DROWN or CHOKE may be either aggressive or merely tragic depending on whether the participant intends them as transitive or intransitive—one can be choked by an assailant, but one can also choke on a pretzel. These ambiguities may also be a source of error variance.

I examine the largest plausible effect size in this task under time pressure by randomly assigning members of a large undergraduate survey class to generate as many or as few aggressive words as possible given the supplied stems.

3.1. Method

Participants were 73 students in an undergraduate lecture course participating as a classroom activity. I did not perform a power analysis; instead, the sample size was set by the number of students in attendance that day. A post-hoc sensitivity analysis conducted after data collection indicated 80% one-tailed power to detect differences between conditions as small as $\delta = 0.73$ (G*Power 3.1, independent samples t-test). Participants performed the word completion task under one of three conditions. In the default condition, participants received the typical task instructions to generate as many words as possible. In the aggressive condition, participants were instructed to generate as many aggressive words as possible. In the nonaggressive condition, participants were instructed to generate as many nonaggressive words as possible. The word list was taken from C. A. Anderson (1999b). Half the words have potential aggressive completions. All participants were given five minutes to complete the measure and were encouraged to skip to the next word stem if a suitable completion did not come to mind. Research assistants later coded the responses using the scoring rubric from C. A. Anderson (1999b), counting up the total number of aggressive word completions provided by each participant.

3.2. Results

On average, without specific instructions, participants generated 6.76 (SD = 2.55) aggressive words. Participants instructed to generate only aggressive words generated 14.71 (SD = 3.96) aggressive words, an increase of $d = 2.36$ over the no-instruction control. Participants instructed to generate only nonaggressive words generated 2.08 (SD = 2.57) aggressive words, a decrease of $d = -1.80$ from the no-instruction control. The two instruction conditions differed by $d = 3.72$.

3.3. Systematic review

I searched PsycINFO, PsycARTICLES, Web of Science, and Scopus using the terms *aggress** AND *word completion task* and *aggress** AND *word fragment*. 31 records were discarded as duplicates. 35 were excluded as irrelevant; several were the similar death-thought accessibility task frequently used in terror management research. Two were excluded for insufficient detail. In the end, 30 unique articles containing 33 studies and 41 effect sizes were retained for inclusion and coding. Effect sizes were extracted and converted to Hedges' g. The distribution of effect size estimates is displayed in Fig. 3.

The average absolute value of the effect size was $g = 0.44$, 95% CI: [0.33, 0.55], $I^2 = 71.40$, $\tau = 0.28$. This average effect size was 24.63% as large as the difference between the participants instructed to generate as many aggressive words as possible and the control condition. It was 18.77% as large as the difference between participants instructed to generate as many and as few aggressive words as possible. No individual effect sizes were noticed as approaching the size of the contrasts found in the present study. Thus, this search does not find any studies reporting implausibly large effects on the word completion task.

Although this study did not find any results of implausible size, it has a few parallels to Study 1. Effect sizes are roughly similar across studies, with an aggression-increasing positive control causing an increase of roughly $d = 1.5$ relative to control and roughly $d = 2.5$ relative to an aggression-decreasing positive control. It also finds that reported effects retrieved through systematic review are, on average, about ~20–25% as large as effects from a maximal positive control. These are, of course, only two studies; further research would be necessary to estimate the average ratio between reported effect sizes and effect sizes from maximal positive controls. Study 2 also demonstrates that, even with explicit instruction to maximize or minimize one's score, there is a moderate amount of variance within conditions.

Both Study 1 and Study 2 have concerned between-subjects effect size estimates on a projective task. In Study 3, I extend the method to a within-subjects effect in a reaction-time task: the aggressive-emotion Stroop task.

4. Study 3: The aggressive-emotion stroop

Most readers will be familiar with the classic version of the Stroop task, in which participants indicate the color of the ink in which text is printed. When the text of the word matches the color of the ink (e.g., the word "red" printed in red), processing either the text or the ink color will prepare the correct response, and reaction times are quick. When the text of the word does not match the color of the ink (e.g., the word "green" printed in red), the prepotent tendency to process the text interferes with processing of the color and preparation of the correct response, and reaction times are slowed.

In contrast to the color Stroop, which measures interference of color words on the ability to indicate a color, the aggressive-emotion Stroop measures interference of emotionally-valenced words on the ability to indicate a color. Participants again indicate the color of the ink, but some words are emotionally neutral while others are aggressive. This does not create an obvious response conflict like the color-word Stroop, but it is theorized that, to the extent that a participant experiences the activation or priming of aggressive thoughts, the tendency to read the word will interfere with naming the word's color. Thus, to the extent that a participant is thinking aggressive thoughts, a greater latency is expected on aggressive-word trials compared to non-aggressive-word trials (Anderson, Anderson, & Deuser, 1996).

This process is more subtle, and researchers tend to find aggression Stroop effects an order of magnitude or two smaller than in the classic Stroop task. Whereas the classic color Stroop finds effects in the range of 65-150 ms (Davidson, Zacks, & Williams, 2003; Haaf & Rouder, 2017), aggressive-emotion Stroop effects are frequently in the tens, and sometimes single digits, of milliseconds. Nevertheless, some studies have reported aggression Stroop effects in the hundreds of milliseconds—effects comparable to, and sometimes even stronger than, the classic color Stroop.

To test the largest plausible aggression Stroop effect, I developed a

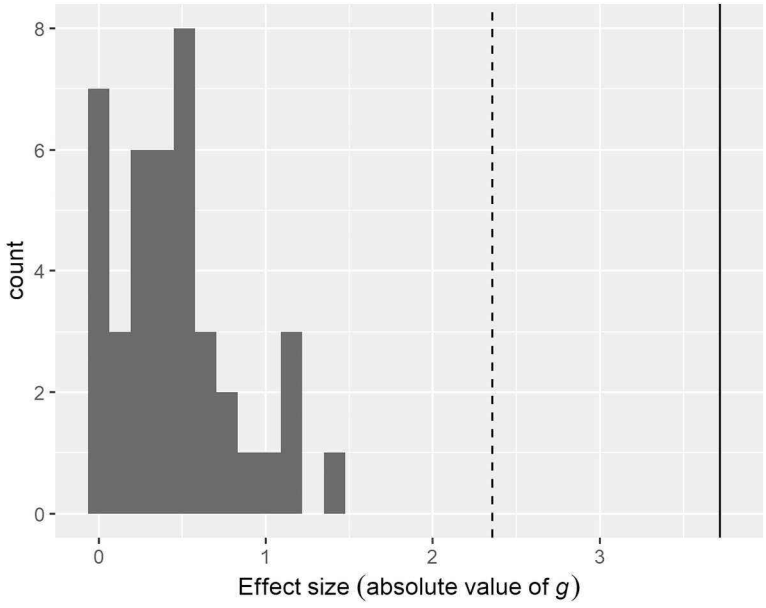


Fig. 3. Effect sizes reported in the word completion task. No effects approach the effect sizes seen in maximal positive control. The dashed vertical line represents the contrast between participants instructed to generate as many aggressive words as possible and participants given no special instructions. The solid vertical line represents the contrast between participants instructed to generate as many aggressive words as possible and participants instructed to generate as few aggressive words as possible.

modified Stroop task in which participants *must both* read the word *and* indicate its color. Thus, in this condition, there is little variance in the degree to which participants read the target word instead of indicating its color: all participants are *required* to fully read the word in addition to identifying the color of the ink.

This condition is compared to a task in which the participant indicates the color of a string of six Xs. In this condition, there is no word to read, and so word-reading cannot interfere with the identification of the color at all. In this way, we can assess the largest plausible difference in reaction times caused by the interference of reading a word's semantic content while identifying the color of its ink.

4.1. Method

The preregistered target sample size was set at 80 because it seemed, heuristically, an appropriate size to get a decent degree of precision in estimation. Eighty-two participants were recruited from Prolific.co, provided informed consent via a consent form on Qualtrics, and performed a PsychoJS computer task on Pavlovia.org.[1] Six participants were excluded for insufficient accuracy, leaving a final sample of 76. Participants were 42 males and 34 females with mean age 31.19 years ($SD = 9.47$) and three not reporting an age. Prolific did not provide data on the ethnic or racial representation of the sample. The task lasted approximately 15 min, and participants were paid $1.65 for their time. This study was preregistered at https://osf.io/hfvsk/. All participants provided informed consent (Illinois State University ethics approval IRB-2020-228).

Participants performed two tasks in counterbalanced order. Participants performed both tasks with the left hand placed on the S, D, and F keys and the right hand placed on the J, K, and L keys.

In the complex task condition, participants pressed one of six keys to indicate both the color and content of a word stimulus. For an nonaggressive word, participants used their left hand, pressing S for red ink, D for green ink, and F for blue ink. For an aggressive word, participants used their right hand, pressing J for red ink, K for green ink, and L for blue ink. A fixation cross was presented between trials for a duration of 550-600 ms. Each of twenty aggressive and twenty nonaggressive words were presented once in each color for a total of 120 trials. See supplementary fig. S1 for a schematic.

In the simple task condition, participants pressed one of three keys to indicate the color of a string of Xs. To keep the tasks comparable, this task was also performed with two hands: some trials required participants to use their left hand, and other trials required participants to use their right hand. To indicate which hand to use, participants were first presented a cue of a left or right arrow. Participants were instructed to respond to this cue by pressing any of the three buttons on that hand. Once the participant pressed a button with the cued hand, they were then presented the target string of six Xs in red, green, or blue ink. Participants would then press the appropriate key using the cued hand, on the left hand pressing S for red, D for green, and F for blue, and on the right hand pressing J for red, K for green, and L for blue. After the participant's response to the target, a fixation cross was presented for 250-300 ms before the next arrow cue. Each cue was presented with each color twenty times each for a total of 120 trials. See supplementary fig. S2 for a schematic.[2]

Participants performed four practice blocks. First, participants practiced using the left hand to indicate the color of nonaggressive words (15 trials). Second, they practiced using the right hand to indicate the color of nonaggressive words (15 trials). Third, they practiced using the left and right hands to sort aggressive and nonaggressive words (20 trials). Lastly, they practiced using both hands to sort words according to both their aggressive content and the color of the ink (30 trials). To reduce errors and latency caused by forgetting the response mapping, the response mapping was displayed on the screen throughout.

To be included in the analysis, participants had to perform at 50% accuracy or greater on both tasks. (Chance accuracy would be 17%.) For each participant, reaction times for correct trials were averaged within each task and a difference score calculated.

Two small changes were made from the preregistration. First, the preregistration indicated that participants would be retained if they had ≥50% accuracy across *all* trials; this was changed to be a requirement on *each* task in case participants misunderstood or withheld effort on one task but not the other. Second, trials with latency in excess of 10s were

[1] Two inattentive participants were identified, excluded, and replaced during the batch data collection, yielding 82 total participants instead of the preregistered 80.

[2] Button mappings could not be counterbalanced across participants due to limitations of the PsychoJS program used for online data collection. Because the measure of interest is a within-subjects difference not confounded with button mapping, this is unlikely to be a major limitation of the study.

discarded as outliers.

4.2. Results

As expected, responses were slower in the complex task, $M = 1235$ ms ($SD = 349$ ms), than in the simple task, $M = 823$ ms ($SD = 221$ ms). Across participants, the average difference score was 412 ms ($SD = 289$ ms), 95% CI [346 ms, 477 ms]. This is a large difference in reaction times, about 2–4 times larger than the color Stroop effect. However, it is interesting to know that even this obvious effect does not exceed half a second in latency.

4.3. Systematic review

I searched PsycINFO, PsycARTICLES, Web of Science, and Scopus using the terms *aggress* emotion** and *Stroop*. 67 records were discarded as duplicates. 79 were excluded as irrelevant: typical reasons for exclusion at this stage included using a different measure (e.g., the typical color Stroop or an emotion Stroop not related to aggression) or research design. Data from seven studies could not be coded due to insufficient detail in reporting. Twenty-seven studies were retained for inclusion and coding, reporting emotion-Stroop effects in 88 cells. Reported aggressive-emotion Stroop effects in ms were extracted from each cell of each study. The distribution of aggressive-emotion Stroop effects is presented in Fig. 4.

Some studies reported effect sizes equal to or greater than the effect observed in maximal positive control. Smeijers, Bulten, Buitelaar, and Verkes (2018) reported emotion Stroop effects ranging from −15.5 s to +1.3 s. Sun, Wang, and Bai (2019) reported emotion Stroop effects of up to 400 ms. Two other studies reported emotion-Stroop effects comparable to the color Stroop effect but smaller than the maximal positive control (Sani, Tabibi, Fadardi, & Stavrinos, 2017; Zhang et al., 2019).

Meta-analysis was not possible due to the infrequency with which standard errors of the difference score were reported. The median absolute value of the aggressive-emotion Stroop effect was 16.5 ms—about 4% as large as the maximal positive control. Among the studies reviewed here, some cells of some experiments reported effects 45% (Zhang et al., 2019), 98% (Sun et al., 2019), and 3760% (Smeijers et al., 2018) as large as the maximal positive control. It may be beneficial to double-check the reports containing these effect sizes.

Thus, the approach of maximal positive controls can be applied to both between-subjects and within-subjects effects and both tasks involving the coding of responses and the calculation of reaction time differences. In this third study, only one study was found to report effects fully in excess of the effect from maximal positive control (Smeijers et al., 2018). An inspection of that paper's methods indicates that this effect is impossible—the task is described as having a reaction-time deadline of 1500 ms, making it impossible to have a difference between conditions of 15,000 ms. Another study was found to approach, but not to exceed, the effect size from maximal positive control (Sun et al., 2019). Although it may be rare to find effect sizes completely exceeding those found in maximal positive control, the combination of maximal positive controls and systematic review may be helpful in evaluating the plausibility of certain effect sizes.

In contrast to Studies 1 and 2, the typical effect was much smaller than that found in maximal positive control (here, 4%, compared to 21% in Study 1 and 25% in Study 2). This may be a feature of the particular subtlety of the emotion-Stroop effect, or it may be that within-subject differences are stronger in maximal positive controls than between-subject differences. I encourage other researchers to experiment with maximal positive controls to explore their properties in relation to primary research.

5. General discussion

Maximal positive controls can provide a cost-effective way to establish the upper bound of plausible effect sizes in a measure. These upper bounds can be useful in detecting errors in previously published literature. Although implausibly large effect sizes may indicate errors in data collection, errors in analysis, or even possible misconduct, it has been my experience that journals are reluctant to issue expressions of concern for implausibly large results. This reluctance may be caused by the difficulty in determining which results are "too big"—a subjective decision that depends on the judgments and expectations of individual researchers and editors. These individual judgments may be better aligned through the empirical support provided by the collection of maximal positive controls. In this way, maximal positive controls might help identify erroneous reports by providing an empirical estimate of how big is too big.

The three examples provided here revealed some possibly erroneous reports. Study 1 suggests that even the largest effect sizes observed on the story completion task should nevertheless be smaller than those repeatedly reported by Hasan et al. (2012, 2013, 2015). This indicates some manner of confound or error in the study. Because of this likely error, it is not clear that the inferences from Hasan et al. (2013) are correct: Violent video games might not increase hostile-world beliefs and aggressive behavior, hostile-world beliefs might not mediate effects of violent games on aggressive behavior, and effects of violent video games (if any) might not accumulate from day to day. To my knowledge, the only other such long-term experiment was that of Kühn et al. (2019),

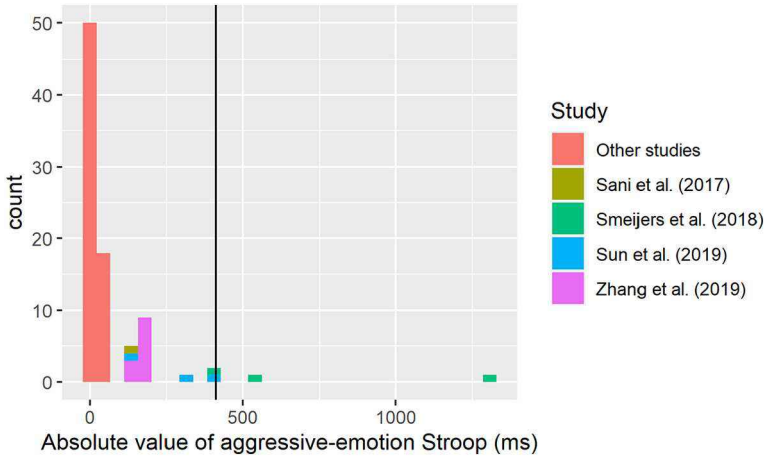


Fig. 4. Histogram of absolute values of aggressive-emotion Stroop scores per cell. The vertical line indicates the effect observed in the maximal positive control. *Note:* One outlier of 15,471 ms (Smeijers et al., 2018) is not included in this graph.

who observed that two months of *Grand Theft Auto V* caused an increase in word completion task scores, but no significant increase on a measure of aggressive world view, an aggressive-cognition lexical decision task, or the Buss-Perry aggression questionnaire. New research will be necessary to test these claims.

Study 3 similarly suggests that even the strongest aggressive-emotion Stroop effect should not exceed about 400 ms. A review of the literature finds a few aggression-emotion Stroop differences of comparable or greater magnitude (Smeijers et al., 2018; Sun et al., 2019). There may be value in double-checking the accuracy of these reports.

In Study 2, by contrast, no studies using the word completion task approached the large effects found using maximal positive controls. Although individual differences in verbal skill may still represent a source of nuisance variance in this task, such differences do not seem to substantially limit the effect sizes one could obtain on this measure.

Researchers using these tasks may benefit from considering the effect size estimates in this study as benchmarks. For example, in the story completion task, if the difference between a peaceful architect and a mass murderer is $d = 2.5$, and the difference between that architect and an extreme sports enthusiast is $d = 1.3$, researchers should expect to find smaller effect sizes when using subtler manipulations and asking about the task's usual generic characters like "Todd" and "Jane." Similarly, in the aggressive emotion Stroop, researchers should expect to find emotion Stroop effects of no more than 400 ms. When researchers estimate how many trials per participant or participants per study they should collect, reference to these estimates may help to inform power analyses by suggesting firm upper limits on even the most optimistic of effect size estimates. In the future, researchers may be able to develop heuristics about the typical ratio between an effect size observed in maximal positive control and in primary research.

One last practical suggestion can be made regarding the administration of the story completion task. Researchers can benefit from considering the influence of the different story stems, which elicited different mean scores. Although it is desirable to use multiple task stimuli to improve the task's generalizability, failing to model the effects of stimulus will leave those effects as error variance, reducing the effect size and degrading study power. The Condition × Scenario interaction suggests that the car accident scenario may be more sensitive than the other scenarios, perhaps by avoiding a floor effect.

Researchers are encouraged to use maximal positive controls to inspect the plausibility of effect sizes reported in their literatures. Maximal positive controls may be collected at lower cost than direct replications. Because maximal positive controls are deliberately dissimilar from original studies, they may also avoid some concerns common to direct replications such as omitted moderators (Stroebe & Strack, 2014), contextual sensitivity of effects (Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016), or the presence or absence of researcher "flair" (Baumeister, 2016). These concerns may avoided when there is a strong logical case that the maximal positive control should yield an effect strictly larger than the original work. Through the use of this method, researchers may learn more about the properties of their measurements, the range of plausible effect sizes, and the quality of research data, thereby facilitating faster scientific self-correction and improving the quality of data used in theory development.

Open practices

Raw data and materials for Studies 1, 2, and 3 and the meta-analyses presented alongside them are available at https://osf.io/7um6d/. Study 3 was successfully preregistered at https://osf.io/hfvsk.

Acknowledgements

I thank Ryan Barry, Taylor Lingle, Sydney Olshak, Joie Pecoraro, Louis Sanchez, and Hannah Westphal for their help in data collection and coding. I also thank Robert Calin-Jageman for inspiring this work through a conversation at SIPS2018, Roger Giner-Sorolla for suggesting the maximal positive control used in Study 1, and Dawn McBride for consultation regarding task design in Study 3.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2020.104082.

References

Akhtar-Danesh, N., & Dehghan-Kooshkghazi, M. (2003). How does correlation structure differ between real and fabricated data-sets? *BMC Medical Research Methodology, 3* (1), 18.

American Psychological Association. (2015). *Resolution on Violent Video Games*.

Anderson, C. A. (1999a). *Story completion task*.

Anderson, C. A. (1999b). *Word completion task*.

Anderson, C. A., Anderson, K. B., & Deuser, W. E. (1996). Examining an affective aggression framework weapon and temperature effects on aggressive thoughts, affect, and attitudes. *Personality and Social Psychology Bulletin, 22*(4), 366–376.

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.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), 1–48. Doi:10.18637/jss.v067.i01.

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology, 66*, 153–158. https://doi.org/10.1016/j.jesp.2016.02.003.

Bushman, B. J. (2016). Violent media and hostile appraisals: A meta-analytic review. *Aggressive Behavior, 42*(6), 605–613.

Bushman, B. J., & Anderson, C. A. (2002). Violent video games and hostile expectations: A test of the general aggression model. *Personality and Social Psychology Bulletin, 28* (12), 1679–1686. https://doi.org/10.1177/014616702237649.

Calvert, S. L., Appelbaum, M., Dodge, K. A., Graham, S., Nagayama Hall, G. C., Hamby, S., … Hedges, L. V. (2017). The American Psychological Association task force assessment of violent video games: Science in the service of public interest. *American Psychologist, 72*(2), 126.

Davidson, D. J., Zacks, R. T., & Williams, C. C. (2003). Stroop interference, practice, and aging. *Aging, Neuropsychology, and Cognition, 10*(2), 85–98.

Greitemeyer, T., & Mügge, D. O. (2014). Video games do affect social outcomes. *Personality and Social Psychology Bulletin, 40*(5), 578–589. https://doi.org/10.1177/0146167213520459.

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in bayesian mixed models. *Psychological Methods, 22*(4), 779.

Hasan, Y., Bègue, L., & Bushman, B. J. (2012). Viewing the world through "blood-red tinted glasses": The hostile expectation bias mediates the link between violent video game exposure and aggression. *Journal of Experimental Social Psychology, 48*(4), 953–956.

Hasan, Y., Bègue, L., & Bushman, B. J. (2015). To be cruel or kind? Hostile expectation bias mediates the link between violent video games and aggression and prosocial behavior. In *International Society for Research on Aggression*.

Hasan, Y., Bègue, L., Scharkow, M., & Bushman, B. J. (2013). The more you play, the more aggressive you become: A long-term experimental study of cumulative violent video game effects on hostile expectations and aggressive behavior. *Journal of Experimental Social Psychology, 49*(2), 224–227.

Hilgard, J. (2019). Comment on Yoon and Vargas (2014): An implausibly large effect from implausibly invariant data. *Psychological Science, 0956797618815434*.

Hilgard, J., Engelhardt, C. R., Rouder, J. N., Segert, I. L., & Bartholow, B. D. (2019). Null effects of game violence, game difficulty, and 2D: 4D digit ratio on aggressive behavior. *Psychological Science, 0956797619829688*.

Kühn, S., Kugler, D. T., Schmalen, K., Weichenberger, M., Witt, C., & Gallinat, J. (2019). Does playing violent video games cause aggression? A longitudinal intervention study. *Molecular Psychiatry, 24*, 1220–1234. https://doi.org/10.1038/s41380-018-0031-7.

Levelt. (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel*.

Moery, E., & Calin-Jageman, R. J. (2016). Direct and conceptual replications of Eskine (2013) organic food exposure has little to no effect on moral judgments and prosocial behavior. *Social Psychological and Personality Science, 7*(4), 312–319.

Pelosi, A. J., & Appleby, L. (1992). Psychological influences on cancer and ischaemic heart disease. *BMJ [British Medical Journal], 304*(6837), 1295–1298.

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331.

Rozin, P. (2006). Naturalness judgments by lay Americans: Process dominates content in judgments of food or water acceptability and naturalness. *Judgment and Decision making, 1*(2), 7.

Sani, S. R. H., Tabibi, Z., Fadardi, J. S., & Stavrinos, D. (2017). Aggression, emotional self-regulation, attentional bias, and cognitive inhibition predict risky driving behavior. *Accident Analysis & Prevention, 109*, 78–88. https://doi.org/10.1016/j.aap.2017.10.006.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). Life after P-Hacking. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2205186.

Smeijers, D., Bulten, E., Buitelaar, J., & Verkes, R.-J. (2018). Associations between neurocognitive characteristics, treatment outcome, and dropout among aggressive forensic psychiatric outpatients. *International Journal of Offender Therapy and Comparative Criminology, 62*(12), 3853–3872. https://doi.org/10.1177/0306624X17750340.

Sosso, F. A. E., Kuss, D. J., Vandelanotte, C., Jasso-Medrano, J. L., Husain, M. E., Curcio, G., … Toth, A. J. (2020). RETRACTED ARTICLE: Insomnia, sleepiness, anxiety and depression among different types of gamers in African countries. *Scientific Reports, 10*(1), 1937. https://doi.org/10.1038/s41598-020-58462-0.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*(1), 59–71. https://doi.org/10.1177/1745691613514450.

Sun, L.-R., Wang, P., & Bai, Y.-H. (2019). Effect of implicit prejudice on intergroup conflict: The cognitive processing Bias perspective. *Journal of Interpersonal Violence*. https://doi.org/10.1177/0886260519844271, 0886260519844271.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences, 113*(23), 6454–6459.

Yoon, G., & Vargas, P. T. (2014). Know thy avatar. *Psychological Science, 25*(4), 1043–1045. https://doi.org/10.1177/0956797613519271.

Zhang, Q., Cao, Y., Gao, J., Yang, X., Rost, D. H., Cheng, G., … Espelage, D. L. (2019). Effects of cartoon violence on aggressive thoughts and aggressive behaviors. *Aggressive Behavior*. https://doi.org/10.1002/ab.21836. ab.21836.

Hartgerink, C. H., Voelkel, J. G., Wicherts, J. M., & van Assen, M. A. L. M. (2019). Detection of data fabrication using statistical tools. Doi:10.31234/osf.io/jkws4.