p-Curve Analysis of the Taylor Aggression Paradigm: Estimating Evidentiary Value and Statistical Power Across 50 Years of Research

Samuel J. West[1*], Courtland S. Hyatt[2], Joshua D. Miller[2], David S. Chester[1]

[1]Department of Psychology, Virginia Commonwealth University, USA

[2]Department of Psychology, University of Georgia, USA

in press at *Aggressive Behavior*

*Correspondence should be addressed to Samuel J. West, 806 W. Franklin St., Virginia Commonwealth University, Richmond, VA 23284, USA; westsj3@vcu.edu

Acknowledgements

We thank Uri Simonsohn for providing his valuable guidance during the conduct of this project.

Disclosure Statement

The authors of this manuscript declare no conflicts of interest regarding this project.

Abstract

The overall reliability or *evidentiary value* of any body of literature is established in part by ruling out publication bias for any observed effects. Questionable research practices have potentially undermined the evidentiary value of commonly-used research paradigms in psychological science. Subsequently, the evidentiary value of these common methodologies remains uncertain. To quantify the severity of these issues in the literature, we selected the Taylor Aggression Paradigm (TAP) as a case study and submitted 170 hypothesis-tests spanning over 50 years of research to a preregistered *p*-curve analysis. The TAP literature (N = 24,685) demonstrated significant evidentiary value but yielded a small average effect size (d = .29) and inadequate power (38%). Main effects demonstrated greater evidentiary value, power, and effect sizes than interactions. Studies that tested the effects of measured traits did not differ in evidentiary value or power to those that tested the effects of experimentally-manipulated states. Exploratory analyses revealed that evidentiary value, statistical power, and effect sizes have improved over time. We provide recommendations for researchers who seek to maximize the evidentiary value of their psychological measures.

Keywords: *p*-curve, evidentiary value, meta-analysis, Taylor Aggression Paradigm, aggressive behavior

Introduction

Methods once considered to be rigorous by psychological scientists have fallen under much needed scrutiny over the last decade (e.g., Simmons, Nelson, & Simonsohn, 2011). Common methodological practices, design considerations, and systemic factors have been identified as contributing to the so-called 'replication crisis' (Ioannidis, 2005; Nosek, Spies, & Motyl, 2012). This crisis has been fueled, in part, by undisclosed and unjustified flexibility in scoring and analysis (i.e., 'researcher degrees-of-freedom') — examples include testing a hypothesis across different scoring strategies and adding or removing covariates until statistical significance is reached (Wicherts et al., 2016). Researcher degrees-of-freedom are couched within a broader category: 'questionable research practices', alongside the use of underpowered samples and the presentation of exploratory analyses as confirmatory (Bakker, van Dijk, & Wicherts, 2012; Ioannidis, 2005; Kerr, 1998). A field-wide prejudice against null results yields an overly optimistic view of the literature, as it renders the number of failed replications of any given effect impossible to know (Greenwald, 1975). Each of these practices undermine psychological science's *evidentiary value* — the extent to which accrued scientific evidence reflects true effects and not false positives.

Evidentiary value is determined by ruling out publication bias as the primary factor for an observed effect in the published literature (Simonsohn, Nelson, & Simmons, 2014a). Multiple investigations have sought to estimate the extent to which prominent findings in psychology possess evidentiary value via direct replications and meta-analytic approaches (Motyl et al., 2017; Open Science Collaboration, 2015; Simonsohn et al., 2014a). However, the evidentiary value of the field's *measures* has received less attention. Research findings are predicated on the evidentiary value of the measures used to obtain them, making it crucial for psychological scientists to establish whether their psychometric approaches hold empirical water. In what follows, we used the *p*-curve (Simonsohn et al., 2014a), to estimate the evidentiary value of a commonly utilized paradigm in psychological science to study an outcome of great importance: the Taylor Aggression Paradigm (TAP; Taylor, 1967).

The Taylor Aggression Paradigm

The TAP is a laboratory protocol in which participants are informed they are engaging in a reaction-time competition against an ostensible opponent (Bushman & Baumeister, 1998; Taylor, 1967). The original TAP publication has been cited by 909 times from 1967 to 2020 (estimates from Google Scholar). At the outset of each trial, participants are given the opportunity to select the intensity and duration of an aversive stimulus (e.g., noise blasts; electric shock) to send to their opponent while the opponent makes the same selection. The intensity, duration, and/or frequency of the aversive stimulus selected by participants serves as the criterion measure of aggressive behavior.

TAP scores are positively associated with personality traits related to an antagonistic interpersonal approach (e.g. psychopathy, narcissism), the frequency of past anti-social behaviors, and are higher in violent offenders than non-offenders (Bettencourt, Talley, Benjamin, & Valentine, 2006; Hammock & Richardson, 1992; Hartmann, 1969; Hyatt, Zeichner, & Miller, 2019; Wolfe & Baron, 1971). Further, TAP scores are positively associated with other behavioral and self-report measures of aggression, increased by provocation, and capture vengeful motivations towards harm-doing (Chester & Lasko, 2018). Demonstrating discriminant validity, TAP scores are not associated with self-reports of competitiveness or pro-social behavior (Bernstein, Richardson, & Hammock, 1987; Giancola & Zeichner, 1995). Across these forms of evidence, the TAP appears to be a valid approach to the laboratory assessment of aggressive behavior.

Critiques of the TAP

Despite the evidence pointing to the TAP as a valid approach to the laboratory measurement of aggressive behavior, it has faced skepticism. Critics have argued that the TAP lacks external validity (Ferguson, 2007) and confounds aggression with competitiveness (Tedeschi & Quigley, 2000; c.f. Chester & Lasko, 2018; Hyatt et al., 2018). Further, the convergent and external validity of the TAP has failed to replicate in some studies (Ferguson & Rudea, 2009; Ferguson, Smith, Miller-Stratton, Fritz, & Heinrich, 2008). The construct validity of the TAP, as with any other measure, is paramount to the integrity of the body of published literature that implements it. Another factor that has implications for the literature surrounding any given task is the evidentiary value of the findings produced by that measure, whereas construct validity is inconsequential if the relevant literature does not exhibit evidentiary value.

The evidentiary value of any given body of literature is determined on the basis of how reliably the effect central to that body of literature is replicated (Simonsohn et al., 2014a). Application of this concept to a given measure stands to provide insight regarding the impact of various methodological practices and research designs common to that measure. For example, critics of the TAP have pointed out that a lack of task standardization may undermine the evidentiary value of this paradigm (Elson et al., 2014; c.f., Hyatt, Chester, Zeichner, & Miller, 2019). The threat to evidentiary value posed by the lack of standardization in implementations of the TAP does not directly impact the construct validity of the task, but may obscure the interpretation of any effects observed using the TAP. The TAP is thus an ideal case study for examining various factors that might impact the evidentiary value of a given task due to strong evidence and arguments both for (e.g., Giancola & Zeichner, 1995) and against (e.g., Ferguson, 2007) its validity, as well as its broad use in psychological science. Evidentiary value in this instance thus provides insight into both the value of the TAP as a measure and how that value may be impacted by certain practices in implementation and analysis.

The statistical power of the TAP literature also remains unknown. Statistical power has been identified by researchers as contributing to the replication crisis in psychology (Abraham & Russell, 2008; Świątkowski & Dompnier, 2017). Though the use of multiple underpowered samples can be a more efficient strategy in achieving a significant result compared to well-powered samples, this practice leads to inflated Type I error rates (Bakker et al., 2012; Ioannidis, 2005). The assessment of the statistical power of research implementing the TAP allows for inferences regarding the status of psychological research at large due to its prevalence.

The Present Research

The aim of the present study was to use the *p*-curve framework to provide estimates of the evidentiary value, statistical power, and effect sizes of the published literature surrounding the TAP. Given the evidence for the validity of the TAP (Bernstein et al., 1987; Chester & Lasko, 2018; Giancola & Zeichner, 1995; c.f. Ferguson, 2007; Ferguson et al., 2008; Ferguson & Rudea, 2009; Tedeschi & Quigley, 2000), and successful replication of various effects using the TAP (e.g., the rejection-aggression link; Chester & DeWall, 2017, Gaertner, Iuzzini, & O'Mara, 2008) we predicted that the TAP literature would exhibit significant evidentiary value. Consistent with critiques of statistical power in psychological science, we also expected that studies using the TAP would be underpowered (Bakker et al., 2012).

We further hypothesized that studies that used the TAP to test main effects would exhibit more evidentiary value and statistical power than those that tested interaction effects. Indeed, with all other parameters held equal, a much larger sample size is needed in order to provide appropriate statistical power to test an interaction than a main effect (Gelman, 2018). Additionally, we hypothesized that studies that used a trait-based measure of an independent variable (e.g., a personality trait questionnaire) would exhibit more evidentiary value and statistical power than those that employed an experimentally-manipulated, state-like independent variable (e.g., an anger induction). We expected this for psychometric reasons — while there are many well-validated trait measures (e.g., NEO PI-R; Costa & McCrae, 2008), experimental manipulations do not often undergo the same rigorous validation process as trait measures, and may even be operationalized as single item, one-off measures. The pregistration of our predictions and methodology (initial registration: https://osf.io/6ekbf; addendum: http://osf.io/7h4qj) along with the data and analysis code (https://osf.io/h9c85/files/) needed to reproduce our findings are publicly available.

Method

p-Curve Analysis

The *p*-curve is a meta-analytic tool that quantifies the evidentiary value and statistical power of any body of published literature (Simonsohn et al., 2014a). The *p*-curve examines evidentiary value by ruling out the likelihood of publication bias for observed effects in a given body of literature. Because the true impact of publication bias cannot be estimated without access to the file drawer, the *p*-curve only examines significant effects. The underlying logic of the *p*-curve relies on characteristics of the *p*-value and null hypothesis significance testing. Specifically, for any true effect the distribution of all statistically-significant *p*-values should be positively and exponentially skewed. Otherwise stated, the majority of reported *p*-values should be closer to .001 than .050 for any true effect in a given body of literature. Conversely, the distribution of *p*-values for any null effect should be flat, such that one has an equal chance of obtaining any given *p*-value. The *p*-curve thus tests the distribution of entered *p*-values for positive-skew and provides statistical power and effect size estimates as computed from the reported degrees of freedom and test statistics (Simonsohn et al., 2014b). When the *p*-curve fails to find positive skew (e.g., a majority of results close to the .05 cutoff) publication bias cannot be ruled out as the reason for a given effect, leaving the evidentiary value of that literature unclear.

The online *p*-curve application (http://www.p-curve.com; Simonsohn et al., 2014a) produces two tests that must be considered jointly before determining the evidentiary value of the statistically-significant hypothesis tests submitted to the analysis – the full curve and the half curve. The full curve tests the distribution of selected *p*-values below .05 for significant right-skew, whereas the half curve tests the right-skew of the distribution of *p*-values below .025 and is more robust against *p*-hacking (Simonsohn, Simmons, & Nelson, 2015). Any set of studies are considered to hold evidentiary value when either the half *p*-curve yields a $p < .05$ right skew test or both full and half curves yield $p < .1$ right-skew tests (Simonsohn et al., 2015).

Literature Search Strategy

Our literature search began by obtaining publications from two existing meta-analyses examining the effects of provocation and various personality traits on laboratory aggression (Bettencourt et al., 2006; Hyatt et al., 2019). These meta-analyses were selected because they were the most extensive meta-analytic reviews of laboratory measures of behavioral aggression currently available. Further, the meta-analysis by Bettencourt and colleagues has been highly influential in the literature, as it has been cited 689 times (estimates from Google Scholar). We then reviewed all publications that cited the original validation of the TAP (i.e., Taylor, 1967) and the widely-used computerized version of the TAP, known as the Competitive Reaction Time

Task (Bushman & Baumeister, 1998). We used PsycINFO, PubMed, and Google Scholar to complete our search. We included hypothesis-tests from these sources that met our inclusionary criteria (detailed below). In addition to these sources, we performed an exhaustive review of all search results on Google Scholar using the keywords "Taylor Aggression Paradigm," "Competitive Reaction Time Task" and "noise blast aggression paradigm." Each of these three search terms yielded an insurmountable number of results for a manual search (>1.32 million search results) and were not used to identify articles for inclusion in our analyses.

Inclusionary Criteria

From the candidate studies identified in the literature search, we included studies that (A) utilized the TAP as a dependent measure, (B) included a specific prediction regarding aggressive behavior in the Introduction section of the manuscript, and (C) provided the appropriate inferential statistics from the test of that hypothesis (or information needed to estimate these statistics). The assumptions of the *p*-curve also imposed restrictions on the studies we could include. Specifically, the *p*-curve requires that all test statistics entered into the analysis are independent from other tests (i.e., not derived from the same samples) and must test a stated hypothesis. Further, the *p*-curve only includes significant (p < .05) effects in the analysis. Manuscripts had to be available in the English language. Hypothesis tests had to be two-tailed (as recommended by Ulrich & Miller, 2018). Individual studies from multi-study publications were included as independent entries into the analysis, provided they met the inclusionary criteria. Whether a study met these criteria was not always clear (e.g., studies with vague hypotheses, 'hypothesis tests' that did not appear to correspond to any stated hypothesis). In these cases, ambiguities regarding our inclusion criteria were resolved via a voting procedure. In this voting procedure, two of the authors (one junior and one senior; one from each institution) voted on whether to include or exclude each study. Any tie votes were then subjected to a tie-breaking vote from a third, senior author. Fifty-four publications were subjected to this procedure which ultimately resulted in the inclusion of 15 publications. Reasoning for all excluded studies that appeared to utilize some form of the TAP can be found in our exclusion table (https://osf.io/kgn9h/).

Several TAP variants were too divergent from canonical implementations of the TAP to be included in our *p*-curve. Such variants were modified beyond the common practice of adjusting the number of trials or the type of aversive stimuli, rendering it problematic to equate with the canonical implementations of the TAP. These excluded variants included studies that interspersed TAP trials with those of other tasks (e.g., the Stroop task), set the target of aggression to be the participant themselves (i.e., self-harm), simulated an interaction in which participants imagined that they were a different individual than themselves (e.g., a fictional character), utilized a competition element not related to reaction times (e.g., word recall), or used point subtractions as the dependent measure of aggression. Determinations regarding whether a study deviated too far from the canonical TAP were made via the above voting procedure. The five most common reasons for exclusion are presented in Table 1 along with their respective frequencies.

Table 1

Five most common reasons for exclusion.

Reason	**Frequency**
Used a non-canonical TAP variant	98
Did not make explicit hypotheses regarding the TAP	58
Hypothesized effect was not significant	54
Necessary data/statistic not presented	49
Could not locate copy in English	22

Test Selection

The appropriate test statistics were selected from each study according to the guidelines detailed in the *p*-curve user guide (p-curve.com/guide.pdf). If no official guidance was offered from the *p*-curve guide regarding what test statistic to select for a particular hypothesis test, we selected the statistic that best represented the test of the selected hypothesis. When necessary, we computed missing inferential statistics from the available descriptive values. For example, if the difference between two groups was hypothesized but the *t*-test statistic was not reported, group means, sample sizes, and standard deviations were then used to compute a *t*-score. In the case of multiple tests of a single hypothesis, or multiple hypotheses using TAP scores as the outcome, we selected the first significant hypothesis test presented in the manuscript, as has been done in previously published *p*-curves (e.g., Vadillo, Gold, & Osman, 2016).

The literature search yielded a total of 5,461 publications. Of these publications, 518 unique publications utilized the TAP or some variant of the TAP in at least one study. Of the 518 candidates, 159 publications contained at least one study that met all the inclusion criteria detailed above. Given that several of these publications contained multiple studies that met inclusion criteria, a total of 171 effects were included in our final sample. These studies were then coded and entered into our disclosure table (https://osf.io/h4gyw/). Inferential statistics from these 171 effects were then entered into the *p*-curve application. Comparisons between *p*-curves were conducted by computing the absolute difference of the *Z*-values between each *p*-curve and finding the corresponding significance level for the residual *Z*-value for the full and half curves, respectively (i.e., Z_{resid} ; U. Simonsohn, personal communication, August, 30, 2018).

Results

Sample Characteristics

Our sample of studies includes analysis results comprised of data 24,685 participants. The average age of participants in our sample was 19.22, though multiple studies did not report standard deviation statistics and we were thus unable to estimate the overall standard deviation of participant age. Approximately 42.47% of participants in all samples were female. The vast majority of studies in our sample ($k = 162$) also made some indication of where their samples came from. Of these, 128 were entirely undergraduate college student samples, 29 were community samples, 4 were samples of children, and 1 was an MTurk sample. Only 83 studies from our full sample reported any race statistics. Of these studies, 17 had a minority (less than 50% of the full sample) of white participants, whereas 72.79%[1] of all participants in studies with race data were white.

Evidentiary Value and Power of Studies Utilizing the TAP

We entered all 171 effects into the omnibus *p*-curve analysis, but one of these effects was misreported as significant and was discarded by the *p*-curve. The omnibus *p*-curve analysis indicated significant right-skew of *p*-values on the full, $Z = -9.09$, $p < .001$, and half, $Z = -9.82$, $p < .001$ curves, indicating significant evidentiary value (Figure 1). We would need to remove the 32 lowest *p*-values in our sample before the TAP no longer exhibits evidentiary value. The power estimate for the literature was 38%, 90% CI = 29%, 48%, with a small overall effect size[2], $d = .29$.

[1] One study contained a sample of 3,000 Chinese participants which heavily skewed estimates of racial diversity among participants and was thus excluded for the purposes of this estimate (Zhang et al., 2019).

[2] Three main effects (one trait-like predictor and two state-like predictors), four interactions, and one mediation were excluded from this estimate as they did not contain the information required.

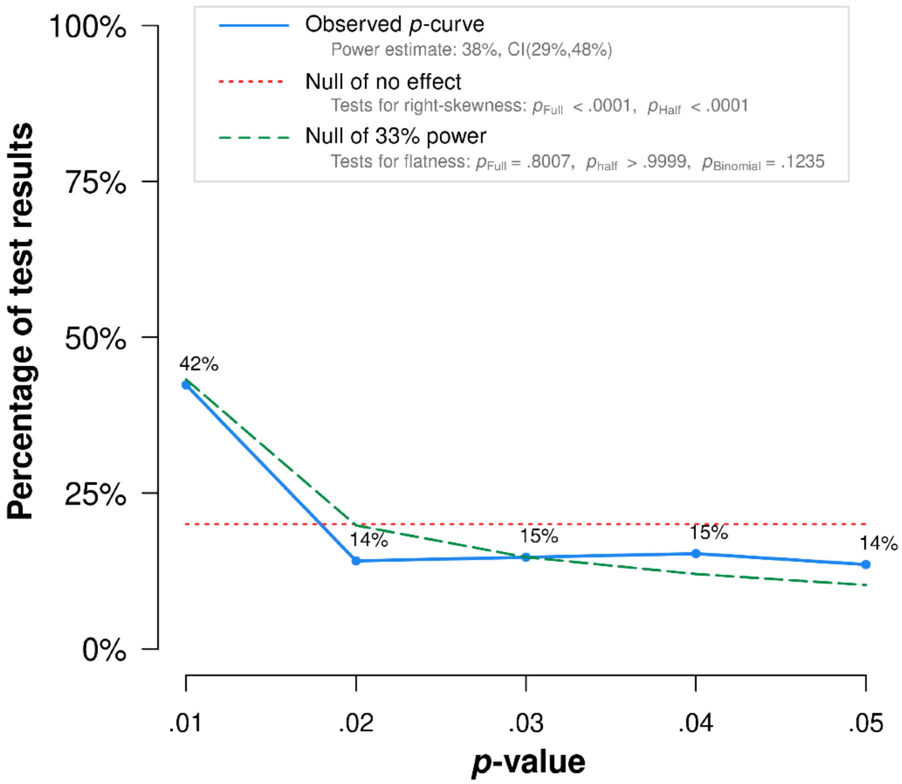

Figure 1. Distribution of *p*-values included in the omnibus *p*-curve analysis including the distribution of *p*-values in the case of no effect and for an underpowered true effect.

Comparing Main Effects and Interactions

Of the selected statistics, 97 were main effects, 70 were interactions, and the remaining three effects were neither a main effect nor an interaction (e.g., mediated-moderation). The main effects *p*-curve (Figure 2A) demonstrated significant right-skew for the full and half curves and thus evidentiary value (Table 2). The interactions *p*-curve (Figure 2B) also demonstrated evidentiary value, as the analysis yielded significant right-skew for the full and half curves. The selected main effects demonstrated significantly more right-skew on both the full, Z_{resid} = -7.68, p < .001, and half *p*-curves, Z_{resid} = -8.30, p < .001, consistent with our prediction that main effects would demonstrate greater evidentiary value than interaction terms. Main effects also demonstrated significantly more statistical power than interactions as their confidence intervals did not overlap, further supporting our expectations. Likewise, the effect size estimate for all main effects was more than three times that of interactions.

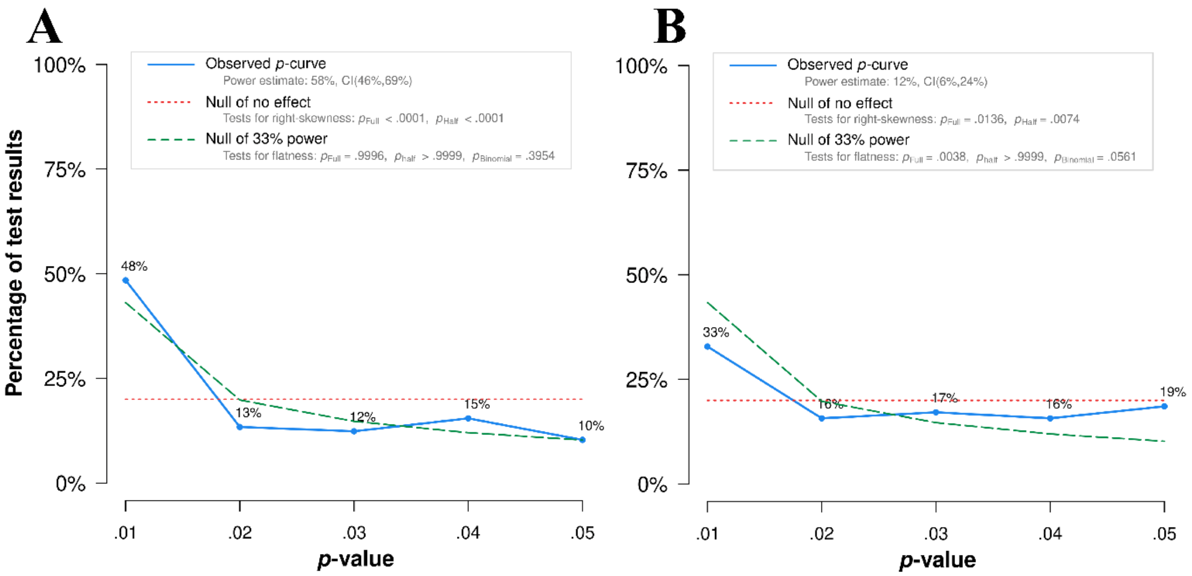

Figure 2. Distribution of *p*-values for all main effects (A) and interactions (B) including the distribution of *p*-values in the case of no effect and for an underpowered true effect.

Table 2

Continuous Test Results and Effect Size Estimates from the Main Effect and Interaction p-Curves.

p-Curve	Full Curve *Z*	Half Curve *Z*	Power	90% CI	*d*
Main Effects	-9.89***	-10.74***	58%	46%, 69%	.37
Interactions	-2.21*	-2.44**	12%	6%, 24%	.10

Notes. CI = confidence interval.

* $p < .05$; **$p < .01$; ***$p < .001$

Comparing Trait and State Independent Variables

Of the selected statistics, 53 utilized measured, trait-like predictors and 44 used manipulated, state-like predictors. The trait predictors (Figure 3A) and state predictors (Figure 3B) yielded significant right-skew for the full and half curves, indicating evidentiary value for both trait-like and state-like predictors of aggression (Table 3). Mixed evidence was found regarding our hypothesis that measured, trait-like predictors would show greater evidentiary value than manipulated, state-like predictors. As trait-like predictors demonstrated significantly more right-skew for the full, Z_{resid} = -2.24, p = .026, but not half p-curve, Z_{resid} = 0.05, p = .960. Inconsistent with our predictions regarding differences in power, the 90% confidence intervals around the power estimates for the trait and state predictors overlapped, indicating no significant difference in statistical power between them.

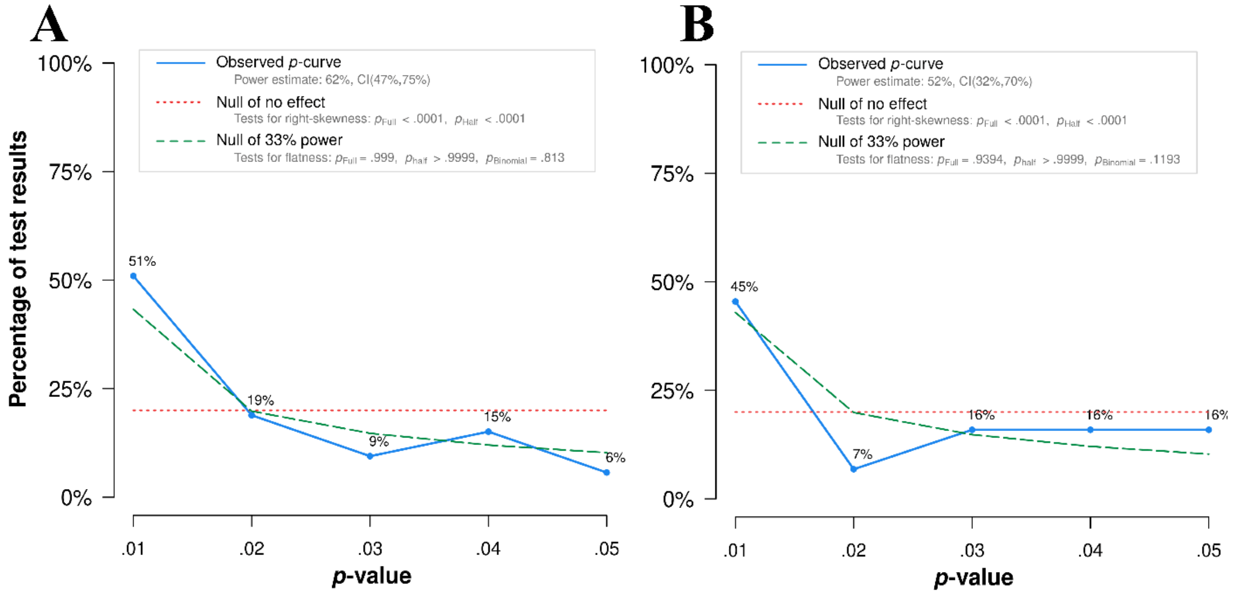

Figure 3. Distribution of p-values for all measured/trait-like (A) and manipulated/state-like predictors (B) including the distribution of p-values in the case of no effect and for an underpowered true effect.

Table 3

Continuous Test Results and Effect Size Estimates from Measured IV and Manipulated IV p-Curves.

p-Curve	Full Curve *Z*	Half Curve *Z*	Power	90% CI	*d*
Trait IVs	-8.07***	-7.61***	62%	47%, 75%	.43
State IVs	-5.83***	-7.66***	52%	32%, 70%	.29

Notes. CI = confidence interval.

* *p* < .05; ***p* < .01; ****p* < .001

Our preregistration plan also called for us to compare the *p*-curves of two groups of TAP scoring strategies: aggregate versus non-aggregate. However, we abandoned these analyses as our attempts to precisely define what constituted an aggregate scoring approach (i.e., studies that aggregated across all available TAP data points) were too general and would have included studies that were not truly aggregate scores. For example, studies that utilized slight variations of the aggregate approach (e.g., combining all data except for the first trial) ended up in our 'non-aggregate' sample, whereas those that utilized a single-trial version of the TAP fell within our definition of the aggregate approach. Instead of attempting to provide a *post hoc* redefinition of these categories, we abandoned this set of hypothesis tests.

Exploratory Analyses: Evidentiary Value and Power Over Time

We conducted exploratory analyses comparing evidentiary value and statistical power between the 89 effects published 2010-2020 and the 60 published 2000-2009 to explore the impact of changing practices (e.g., preregistration) on psychological science across time (akin to Motyl et al., 2017). The *p*-curve for studies published 2000-2009 (Figure 4A) and those published 2010-2020 (Figure 4B) exhibited significant right-skew for the half and full curves indicating evidentiary value in both groups of studies (Table 4). Studies published 2010-2020 demonstrated significantly greater right-skew than those published 2000-2009 on the full, Z_{resid} = -5.76, p < .001, and half Z_{resid} = -2.39, p = .017, p-curves and thus, greater evidentiary value. The 2010s study group also demonstrated more than four times the statistical power of the 2000s study group. Further, the respective confidence intervals for these power estimates did not overlap, indicating a significant difference in power. Similarly, the effect size estimate for the 2010s group was more than three times that of the 2000s group.

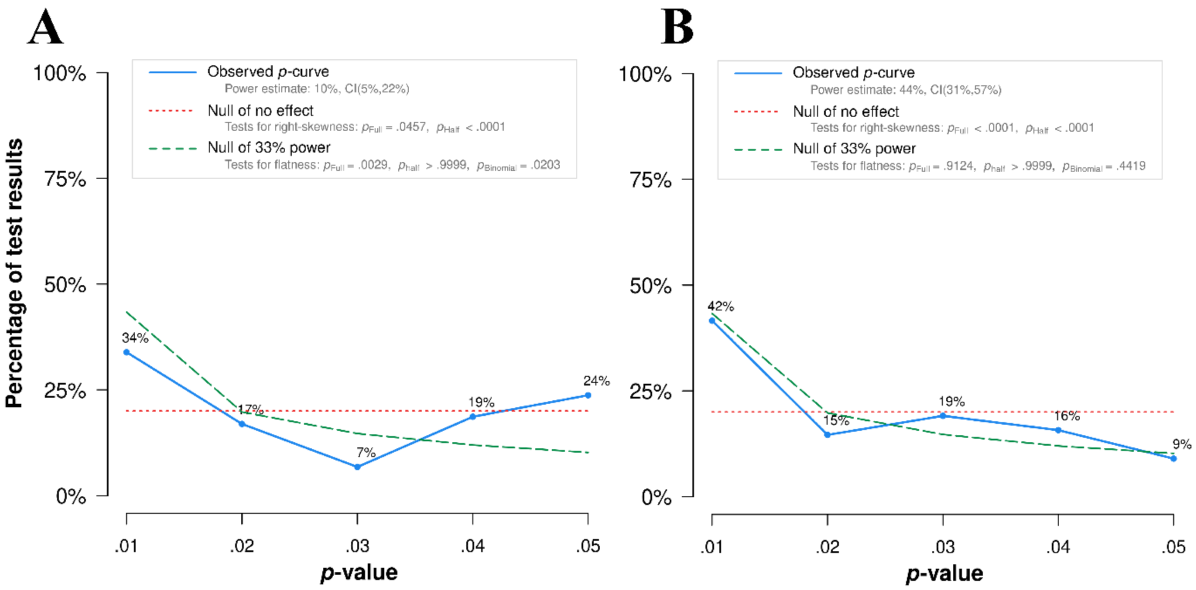

Figure 4. Distribution of p-values for studies published 2000-2009 (A) and those published 2010-2020 (B) including the distribution of p-values in the case of no effect and for an underpowered true effect.

Table 4

Continuous Test Results from p-Curves of Studies Published 2000-2009 and 2010-2018.

p-Curve	Full Curve Z	Half Curve Z	Power	90% CI	d
2000-2009	-1.69*	-3.77***	10%	5%, 22%	.10

2010-2020	-7.45***	-6.16***	44%	31%, 57%	.31

Notes. CI = confidence interval.

* $p < .05$; **$p < .01$; ***$p < .001$

Monte Carlo Simulation

Some have argued that *p*-curve does not handle heterogeneity well (McShane, Böckenholt, & Karsten, 2016; van Aert, Wicherts, & van Assen, 2016). The creators of the *p*-curve have produced simulations under a range of conditions, which indicate that *p*-curve handles heterogeneity well, though these simulations were based on a group of 20 effects (Simmons, Nelson, & Simonsohn, 2018). Thus, we altered the simulation code used by Simonsohn and colleagues to meet the parameters of our current sample of 170 independent effects across 1,000 simulations. Results indicated a 'true' mean power of 38% and a simulated *p*-curve estimate of 42% (Figure 5). Both of the simulation estimates are within the 90% confidence interval of the power estimate from omnibus *p*-curve test.

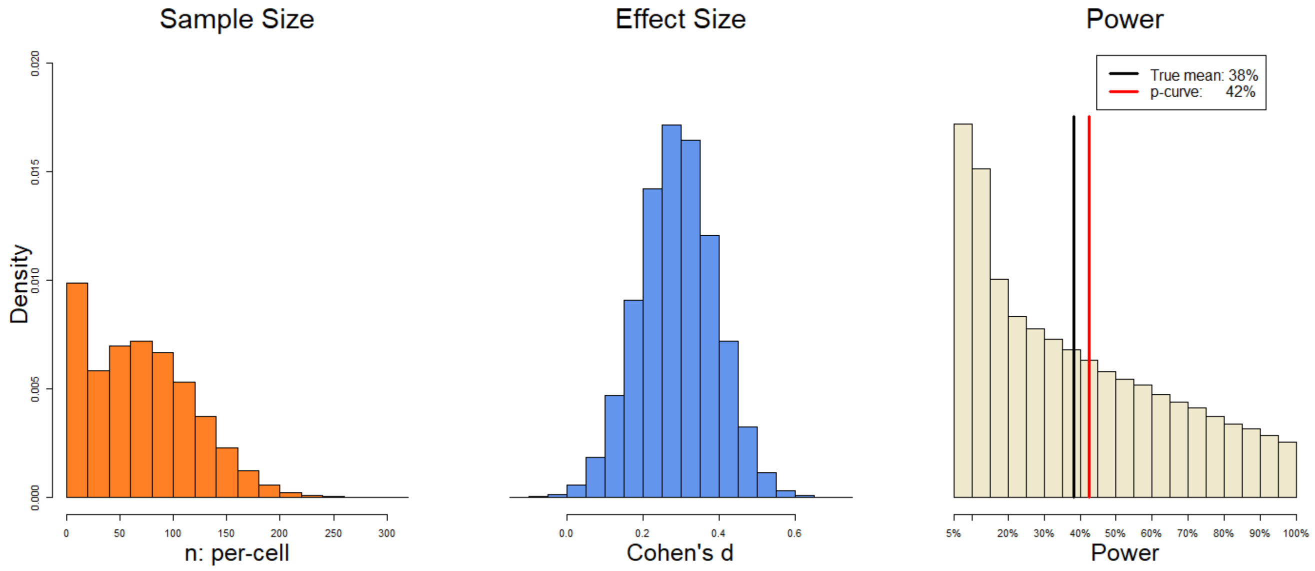

Figure 5. Results of a simulation estimating power across 1000 samples of 170 effects.

Discussion

The evidentiary value of many popular research methods used in psychological science remain uncertain due to questionable research practices (Bakker et al., 2012; Ioannidis, 2005; Nosek et al., 2012; Simmons et al., 2011). The current work presents a novel application of the *p*-curve to the assessment of a common research paradigm (i.e., the Taylor Aggression Paradigm [TAP]; Taylor, 1967), which served as a case study to draw inferences regarding the evidentiary value of behavioral measures of psychological constructs.

Evidentiary Value and Statistical Power of the TAP Literature

Consistent with our preregistered predictions, the TAP demonstrated substantial evidentiary value overall. This finding is bolstered the various effects that have been observed and reliably replicated using the TAP, such as the effect of alcohol on aggression (e.g., Giancola & Corman, 2007; Miller, Parrott, & Giancola, 2009), similarity in the trait predictors of lab-based aggression (Hyatt, et al., 2019), and externalizing behaviors more broadly (Vize, Collison, Miller, & Lynam, 2019). Conversely, this finding is inconsistent with critiques of the TAP pointing to publication bias as an explanation for observed effects (e.g., Ferguson, 2007). As such, the evidence presented in the current work taken in conjunction with the wealth of literature indicating the validity of the TAP and replicability of effects as tested using the TAP, indicate that the TAP indeed holds evidentiary value. We however caution our readers that our analyses do not indicate the evidentiary value of any specific subset of effects that were included in our analyses.

In line with our predictions, the TAP literature was severely underpowered (i.e., 38%). In the most favorable case, our analyses suggest that trait main effects examined using the TAP were powered at 62%. Conversely, tests of interactions were powered at 12%. These findings are consistent with broader trends in psychological science of using underpowered studies to test hypotheses in general (Bakker et al., 2012; Bakker, Hartgerink, Wicherts, & van der Maas, 2016; Szucs & Ioannidis, 2017; Wicherts et al., 2016) and interactions specifically (Gelman, 2018; Kenny, 2015; Nieuwenhuis, Forstmann, & Wagenmakers, 2011). The lack of power found among studies using the TAP is of great concern, as underpowered research is more likely to yield false-positives and false-negatives and less likely to detect true-positives and true-negatives. The effect size estimate of the TAP literature, $d = 0.29$, $r = .14$, was smaller than previous estimates of aggression effect sizes in the psychological literature (e.g., $d = 0.49$, $r = .24$; Richard, Bond, & Stokes-Zoota, 2003). Indeed, a recent meta-analysis found that even the most potent personality predictors of TAP aggression are still relatively modest in size (psychopathy; *mean r* = .23; agreeableness; *mean r* = -.20; Hyatt et al., 2019). These findings should serve as a call for those who use the TAP to increase the power of their investigations — via larger sample sizes, replacing between-participant with within-participant designs, and use of well-validated manipulations.

Our effect size estimates allow future TAP researchers to base *a priori* power estimates on design considerations when no relevant effect sizes exist in the literature for this purpose. Researchers wishing to examine novel phenomenon using the TAP have little guidance in respect to the smallest effect size of interest. Using the smallest effect size of interest in *a priori* power analyses is ideal when no estimates of the expected effect(s) exist in the literature (Albers & Lakens, 2018). As such, TAP researchers can rely on design elements (e.g., trait vs. state IVs) and the complexity of expected effects (e.g., main effects vs. interactions) to determine the needed sample size rather than relying on arbitrary approaches. In general, our findings indicate that irrespective of these elements TAP researchers should plan to power small-to-moderate effects as all of our analyses yielded effect sizes within this range.

Our findings that the TAP literature demonstrates both evidentiary value and poor statistical power, are difficult to reconcile. How could a literature that is largely underpowered exhibit substantial evidentiary value? There are many possible sources of this paradox. First, a minority of studies that were reviewed were adequately powered and these were enough to produce the observed evidentiary value in the literature. Second, the literature may have been testing larger effects with more laboratory precision than accounted for in this *p*-curve analysis — thus increasing statistical power beyond our reported estimates. Third, ambitious *p*-hacking and undisclosed flexibility in implementation and analysis may have exaggerated the underlying evidentiary value in the literature (Elson et al., 2014; Ulrich & Miller, 2015; Wicherts et al., 2016). Similarly, it is also possible that *p*-hacking among the literature included led to artificially deflated power estimates. Future meta-analytic work will need to disentangle these sources and determine whether the TAP literature's evidentiary value is under- or over-stated by our findings.

Effects of Study Design Elements

Consistent with our hypotheses, TAP studies testing main effects demonstrated significantly more evidentiary value and statistical power than studies testing interactions. Further, main effects yielded an effect size estimate more than three-fold that of interactions. As such, investigators should be wary of testing complex interactions with the TAP, unless they are able to adequately power such designs. Further, the robustness of existing interaction findings should be viewed with skepticism until better-powered work is able to establish such criteria.

Contrary to our hypotheses, studies using measured independent variables displayed significantly more evidentiary value than those using manipulations on the full but not half *p*-curve. We placed more weight on the half-curve due to its robustness to *p*-hacking and concluded that these bodies of literature did not differ in evidentiary value. Further, trait measures and state manipulations did not differ in statistical power, though trait measures did yield a nominally greater effect size. Thus, the TAP seems equally capable of capturing both trait and state effects across both correlational and experimental designs. Our findings provide evidence that the TAP is able to capture real effects that arise from both enduring human dispositions, as well as from transient psychological states. A logical corollary of these results is that TAP-measured aggression is therefore a product of both traits and states. This is strong empirical support for one of the core tenets of the General Aggression Model and I^3 meta-theory (Anderson & Bushman, 2002; Finkel & Hall, 2018) — namely, that the level of aggression inherent in any potentially aggressive situation is a product of both dispositional and situational factors. More specific theoretical frameworks of aggression, such as threatened egotism theory (Baumeister, Bushman & Campbell, 2000) and alcohol myopia theory (Steele & Josephs, 1990), which emphasize the interactions between traits and states are thus valid approaches to the study of aggressive acts. Ongoing development of aggression theories should emphasize hypotheses and designs that emphasize the dynamic roles of both traits and states.

Such sensitivity to states and traits is a major boon to any laboratory measurement approach, yet requires that investigators are careful in the inferences they draw using the TAP. Because the TAP is able to capture both state and trait effects, investigators must take steps to ensure that any claims of their effects' generalizability (or lack thereof) across time and situations are supported by the evidence. For instance, if a study reveals a positive association between a state-measure of anger and TAP scores, this may reflect the intended state, or may merely be an artifact of the tendency for dispositionally angry individuals to exhibit greater TAP scores. If the investigators from this study want to infer that this is a state-level effect, then they would need to provide additional evidence that disentangles trait anger from their observed effect. By empirically articulating the traits and states that predict and cause aggression, we can advance our understanding and treatment of such costly harm.

Change Over Time

Our exploratory analyses suggest that the evidentiary value, effect sizes, and statistical power, of studies utilizing the TAP have increased over time. These findings are consistent with recent work demonstrating improved power and evidentiary value across a ten-year period (Motyl et al., 2017) and stand in contrast to findings that imply there have been no improvements in power in the field of psychology (Szucs & Ioannidis, 2017). This is a reason for cautious optimism, as the various initiatives and regulations aimed at creating more reliable, reproducible research have likely had a positive impact on the credibility of our field — though our data cannot speak to the underlying reasons for such improvement.

Limitations

The current work tests the evidentiary value and power of a single methodological paradigm. Although the TAP is widely utilized in a broad body of research spanning over five decades, testing the evidentiary value and power of other measures (e.g., cognitive assessments) may yield different results. We thus caution readers not to use these findings as an ‘all clear’ signal that behavioral measures, much less psychological measures writ large, possess evidentiary value. Instead, we encourage investigators to apply the *p*-curve to their own foundational measures and examine whether the literature built around these tools contain more statistical signal than noise. We also warn our readers that the *p*-curve technique is unable to establish the construct validity of a measure and should not be used as a meta-analytic surrogate for the hard work of validation research. Construct validation is a complex process that aggregates multiple sources of evidence and is a prerequisite for claims of the evidentiary value of a given measure (Cronbach & Meehl, 1955; Flake, Pek, & Hehman, 2017).

Some have argued that the *p*-curve performs poorly under conditions of high heterogeneity, though simulation work has demonstrated that *p*-curve is robust against heterogeneity (McShane et al., 2016; Simmons et al., 2018; van Aert et al., 2016). Our own simulations were consistent with those conducted by Simmons and colleagues (2018) such that there did appear to be some downward bias in power estimates, but this influence was inconsequential to the results of our analyses as both estimates from this simulation were within our 90% confidence interval for the omnibus *p*-curve. Another limitation lies in the nature of our test selection rules, which in some cases led to the selection of very simple, overpowered effects (e.g., main effects of provocation on aggression). However, our results indicated that we would need to drop the 32 lowest *p*-values in our sample (18.82% of our total sample) before the included effects no longer exhibited evidentiary value. Further examination of these values indicated that only 12 of our selected *p*-values were below .0001, where 143 *p*-values were greater than .001, suggesting that our findings are not due to several extremely low *p*-values. Similarly, our inclusion criteria paired with the requirements of the *p*-curve analysis led to the exclusion of the majority of studies found during our search. As such, the studies included in our final sample may not necessarily be reflective of the evidentiary value of the entire body of literature using the TAP. The nature of the *p*-curve does not allow for us to draw inferences regarding these excluded publications beyond documenting their reason for exclusion. We argue however that our sample was indeed representative of the TAP literature, as our sample includes publications spanning multiple areas of psychology, many of which are highly cited and have been quite impactful in the literature. Because the *p*-curve is the only analysis that estimates such evidentiary value, we maintain that our approach and findings are the most representative of the TAP literature as allowed by the constraints of this framework.

Further, the *p*-curve is not able to directly estimate publication bias, the heterogeneity of underlying effects, or the prevalence of *p*-hacking. Without knowing these parameters, researchers are still bound by the file drawer, such that we have no discernable means to estimate how many times a methodological paradigm has failed to reliably produce consistent results across any given body of literature. These unknowns may contribute to the selection of inadequate study designs, measures, and analytical models for any given research question.

Conclusions

The evidence presented here suggests that the published literature using the TAP holds evidentiary value and evinces modestly-sized effects, though these estimates are obscured by underpowered analyses. To the extent that the TAP is representative of laboratory measurement paradigms in psychology, these findings imply that studies using behavioral measures may be assessing signal and not mere noise, but are routinely underpowered. Thus, we strongly encourage psychological scientists to take issues of statistical power seriously such that the overall statistical power of the literature may continue to improve. However, consistent with the proliferation of the 'Open Science Movement' (e.g., Nosek et al., 2015), metrics of best scientific practice seem to be improving over time, suggesting a cautiously optimistic view of the future. We are hopeful that psychological science will continue to build upon on these successes, and continue to move toward encouraging more transparency, preregistration, appropriate statistical power, and methodological rigor — ultimately improving the evidentiary value of our field.

References

Abraham, W. T., & Russell, D. W. (2008). Statistical power analysis in psychological research. *Social and Personality Psychology Compass, 2*, 283–301. doi:10.1111/j.1751-9004.2007.00052.x

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology, 74*, 187-195. doi:10.1016/j.jesp.2017.09.004

Allen, J. J., & Anderson, C. A. (2017). General aggression model. *The International Encyclopedia of Media Effects*, 1-15. doi: 10.1002/9781118783764.wbieme0078

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*, 543-554. doi:10.1177/1745691612459060

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' intuitions about power in psychological research. *Psychological Science, 27*, 1069-1077. doi:10.1177/0956797616647519

Baumeister, R. F., Bushman, B. J., & Campbell, W. K. (2000). Self-esteem, narcissism, and aggression: Does violence result from low self-esteem or from threatened egotism? *Current Directions in Psychological Science, 9*(1), 26-29. doi: 10.1111/1467-8721.00053

Bettencourt, B., Talley, A., Benjamin, A. J., & Valentine, J. (2006). Personality and aggressive behavior under provoking and neutral conditions: a meta-analytic review. *Psychological Bulletin, 132*, 751-777. doi:10.1037/0033-2909.132.5.751

Bernstein, S., Richardson, D., & Hammock, G. (1987). Convergent and discriminant validity of the Taylor and Buss measures of physical aggression. *Aggressive Behavior, 13*, 15-24. doi:10.1016/j.jesp.2011.12.019

Bushman, B. J., & Baumeister, R. F. (1998). Threatened egotism, narcissism, self-esteem, and direct and displaced aggression: Does self-love or self-hate lead to violence? *Journal of Personality and Social Psychology, 75*, 219–229. doi: 10.1037/e413802005-416

Chester, D. S., & Lasko, E. N. (2018). Validating a Standardized Approach to the Taylor Aggression Paradigm. *Social Psychological and Personality Science,* 1-12. doi: 10.1177/1948550618775408

Chester, D. S., & DeWall, C. N. (2017). Combating the sting of rejection with the pleasure of revenge: A new look at how emotion shapes aggression. *Journal of Personality and Social Psychology, 112*, 413-430. doi:10.1037/pspi0000080

Costa Jr, P. T., & McCrae, R. R. (2008). The Revised NEO Personality Inventory (NEO-PI-R). In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment, Vol. 2. Personality measurement and testing* (p. 179–198). Sage Publications, Inc. doi:10.4135/9781849200479.n9

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302. doi: 10.1037/h0040957

Elson, M., Mohseni, M. R., 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, 26*, 419. doi:10.1037/a0035569

Ferguson, C. J. (2007). Evidence for publication bias in video game violence effects literature: A meta-analytic review. *Aggression and Violent behavior*, *12*, 470-482. doi:10.1016/j.avb.2007.01.001

Ferguson, C. J., Smith, S., Miller-Stratton, H., Fritz, S., & Heinrich, E. (2008). Aggression in the laboratory: Problems with the validity of the modified Taylor Competitive Reaction Time Test as a measure of aggression in media violence studies. *Journal of Aggression, Maltreatment & Trauma, 17*, 118–132. doi:10.1080/10926770802250678

Ferguson, C. J., & Rueda, S. M. (2009). Examining the validity of the modified Taylor competitive reaction time test of aggression. *Journal of Experimental Criminology*, *5*, 121. doi:10.1007/s11292-009-9069-5

Finkel, E. J., & Hall, A. N. (2018). The I3 model: A metatheoretical framework for understanding aggression. *Current Opinion in Psychology*, *19*, 125-130. doi: 10.1016/j.copsyc.2017.03.013

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370-378. doi: 10.1177/1948550617693063

Gaertner, L., Iuzzini, J., & O'Mara, E. M. (2008). When rejection by one fosters aggression against many: Multiple-victim aggression as a consequence of social rejection and perceived groupness. *Journal of Experimental Social Psychology, 44*, 958-970. doi:10.1016/j.jesp.2008.02.004

Gelman, A. (2018, March 15). You need 16 times the sample size to estimate an interaction than to estimate a main effect [blog post]. Retrieved from http://andrewgelman.com/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/

Giancola, P. R., & Corman, M. D. (2007). Alcohol and aggression: A test of the attention-allocation model. *Psychological Science, 18*, 649-655. doi:10.1111/j.1467-9280.2007.01953.x

Giancola, P. R., & Zeichner, A. (1995). Construct validity of a competitive reaction-time aggression paradigm. *Aggressive Behavior, 21*, 199–204. doi:10.1002/1098-2337

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1-20. doi: 10.1037/10109-033

Hammock, G. S., & Richardson, D. R. (1992). Predictors of aggressive behavior. *Aggressive Behavior*, *18*, 219-229. doi: 10.1002/1098-2337(1992)18:3%3C219::AID-AB2480180305%3E3.0.CO;2-P

Hartmann, D. P. (1969). Influence of symbolically modeled instrumental aggression and pain cues on aggressive behavior. *Journal of Personality and Social Psychology*, *11*, 280-288. doi:10.1037/h0027071

Hyatt, C. S., Weiss, B. M., Carter, N. T., Zeichner, A., & Miller, J. D. (2018). The relation between narcissism and laboratory aggression is not contingent on environmental cues of competition. *Personality Disorders: Theory, Research, and Treatment*, *9*, 543-552. doi:10.1037/per0000284

Hyatt, C. S., Chester, D. S., Zeichner, A., & Miller, J. D. (2019). Analytic flexibility in laboratory aggression paradigms: Relations with personality traits vary (slightly) by operationalization of aggression. *Aggressive Behavior*, *45*, 377-388. doi: 10.1002/ab.21830

Hyatt, C. S., Zeichner, A., & Miller, J. D. (2019). Laboratory aggression and personality traits: A meta-analytic review. *Psychology of Violence, 9*(6), 675-689. doi:10.1037/vio0000236

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine, 2*, e124. doi:10.1371/journal.pmed.0020124

Kenny, D. A. (2015, March 31). *Moderation.* Retrieved from http://davidakenny.net/cm/moderation.htm

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196-217. doi:10.1207/s15327957pspr0203_4

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science, 11*, 730-749. doi:10.1177/1745691616662243

Miller, C. A., Parrott, D. J., & Giancola, P. R. (2009). Agreeableness and alcohol-related aggression: The mediating effect of trait aggressivity. *Experimental and Clinical Psychopharmacology, 17*, 445-455. doi:10.1037/a0017727

Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., . . . Skitka, L. J. (2017). The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of Personality and Social Psychology, 113*, 34-58. doi:10.1037/pspa0000084

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E. J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience*, *14*, 1105-1107. doi:10.1038/nn.2886

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Contestabile, M. (2015). Promoting an open research culture. *Science*, *348*, 1422-1425. doi:10.1126/science.aab2374

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615-631. doi:10.1177/1745691612459058

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*. doi:10.1126/science.aac4716

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

Simmons J. P., Nelson, L. D., & Simonsohn, U. (2018, January 8). *P*-curve Handles Heterogeneity Just Fine. *Datacolada*. Retrieved from http://datacolada.org/67.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359-1366. doi:10.1177/0956797611417632

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General, 144,* 1146-1152. doi:10.1037/xge0000104

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*, 534-547. doi:10.1037/a0033242

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

Steele, C. M., & Josephs, R. A. (1990). Alcohol myopia: Its prized and dangerous effects. *American Psychologist, 45*(8), 921–933. doi: 10.1037/0003-066X.45.8.921

Świątkowski, W., & Dompnier, B. (2017). Replicability crisis in social psychology: Looking at the past to find new pathways for the future. *International Review of Social Psychology, 30*, 111-124. doi:10.5334/irsp.66

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology, 15*, e2000797. doi:10.1371/journal.pbio.2000797

Taylor, S. P. (1967). Aggressive behavior and physiological arousal as a function of provocation and the tendency to inhibit aggression. *Journal of Personality, 35*, 297–310. doi: 10.1111/j.1467-6494.1967.tb01430.x

Tedeschi, J. T., & Quigley, B. M. (2000). A further comment on the construct validity of laboratory aggression paradigms: A response to Giancola and Chermack. *Aggression and Violent Behavior, 5*, 127-136. doi:10.1016/S1359-1789(98)00028-7

Ulrich, R., & Miller, J. (2018). Some properties of p-curves, with an application to gradual publication bias. *Psychological Methods, 23*, 546-560. doi:10.1037/met0000125

Ulrich, R., & Miller, J. (2015). p-hacking by post hoc selection with multiple opportunities: Detectability by skewness test?: Comment on Simonsohn, Nelson, and Simmons (2014). *Journal of Experimental Psychology: General, 144*, 1137-1145. doi:10.1037/xge0000086

Vadillo, M. A., Gold, N., & Osman, M. (2016). The Bitter Truth About Sugar and Willpower: The Limited Evidential Value of the Glucose Model of Ego Depletion. *Psychological Science, 27*, 1207–1214. doi:10.1177/0956797616654911

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses based on pvalues: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science, 11*, 713-729. doi: 10.1177/1745691616650874

Vize, C. E., Collison, K. L., Miller, J. D**.,** & Lynam, D. R. (2019). Using Bayesian methods to update and expand the meta-analytic evidence of the Five-Factor Model's relation to antisocial behavior. *Clinical Psychology Review, 67,* 61-77. doi:10.1016/j.cpr.2018.09.001

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7*, 1832. doi:10.3389/fpsyg.2016.01832

Wolfe, B. M., & Baron, R. A. (1971). Laboratory aggression related to aggression in naturalistic social situations: Effects of an aggressive model on the behavior of college student and prisoner observers. *Psychonomic Science*, *24*, 193-194. doi:10.3758/BF03335564

Zhang, Q., Cao, Y., Gao, J., Yang, X., Rost, D. H., Cheng, G., Teng, Z., & Espelage, D. L. (2019). Effects of cartoon violence on aggressive thoughts and aggressive behaviors. *Aggressive Behavior, 45*(5), 489–497. doi:10.1002/ab.21836