An empirically based power primer for laboratory aggression research

Courtland S. Hyatt[1]

Michael L. Crowe[2]

Samuel J. West[3]

Colin E. Vize[4]

Nathan T. Carter[1]

David S. Chester[3]

Joshua D. Miller[1]

[1]University of Georgia

[2]VA Boston Healthcare System

[3]Virginia Commonwealth University

[4]University of Pittsburgh

As of 9/16/21, this manuscript is in press at Aggressive Behavior.

Corresponding Author:
Courtland S. Hyatt
University of Georgia
125 Baldwin St., Athens, GA 30602
cshyatt@uga.edu

Funding note: David S. Chester is supported by the NIAAA under award K01AA026647. Colin E. Vize is funded through the National Institute of Mental Health postdoctoral training grant T32MH018269. There are no conflicts of interest to report.

Abstract

Recent reviews suggest that, like much of the psychological literature, research studies using laboratory aggression paradigms tend to be underpowered to reliably locate commonly observed effect sizes (e.g., r = ~.10 to .20, Cohen's d = ~.20 to .40). In an effort to counter this trend, we provide a "power primer" that laboratory aggression researchers can use as a resource when planning studies using this methodology. Using simulation-based power analyses and effect size estimates derived from recent literature reviews, we provide sample size recommendations based on type of research question (e.g., main effect vs. two-way vs. three-way interactions) and correlations among predictors. Results highlight the large number of participants that must be recruited to reach acceptable (~80%) power, especially for tests of interactions where the recommended sample sizes far exceed those typically employed in this literature. These discrepancies are so substantial that we urge laboratory aggression researchers to consider a moratorium on tests of three-way interactions. Although our results use estimates from the laboratory aggression literature, we believe they are generalizable to other lines of research using behavioral tasks, as well as psychological science more broadly. We close by offering a series of best practice recommendations and reiterating long-standing calls for attention to statistical power as a basic element of study planning.

Keywords: power, aggression, research methodology

An empirically based power primer for laboratory aggression research

Aggression is a complex social phenomenon traditionally defined as intentional harm inflicted upon someone who is motivated to avoid this harm (Baron, 1977). This behavior can take many forms, including but not limited to physical aggression, verbal aggression, and relational aggression (i.e., damaging a person's social status), all of which range from milder to more severe presentations. Consistent with this range of manifestations, researchers make use of a myriad of aggression operationalizations. One popular method for studying physical aggression is through laboratory paradigms (Warburton & Bushman, 2019), which generally involve giving participants the opportunity to deliver noxious stimuli (e.g., electric shock, noise blast) to a bogus confederate, often under the guise of a competitive reaction time task.

Laboratory aggression methodology has been criticized on several grounds. For example, critics have raised specific concerns about external validity, such that aggression in the context of the laboratory ostensibly bears little relation to how aggression manifests in "real world" settings (McCarthy et al., 2018; Tedeschi et al., 1996). Additionally, there is a pervasive lack of standardization of analytic strategy in laboratory aggression studies (and psychology more broadly; Simmons et al., 2011) that can impact interpretation of results (Elson et al., 2014; Hyatt et al., 2019). When this latter concern is coupled with a general, field-wide reliance on arbitrary thresholds of statistical significance (i.e., $p < .05$) and a prejudice against null findings (Greenwald, 1975), the result is a publication process that often values exciting (and perhaps non-replicable) results over null findings or more incremental (but perhaps more replicable) contributions (Nosek, et al., 2012).

Statistical Power

Basic issues related to study design and implementation have also been identified as contributing factors to the "replication crisis" in psychology. For example, when designing a study to investigate a hypothesis, how large of a sample does a researcher need to collect to confidently identify an effect size of a certain magnitude? The conventional approach to answering this question in the social sciences has been to conduct a statistical power analysis (Cohen, 1969, 1992), where statistical power refers to the likelihood of correctly rejecting the null hypothesis when the population effect $\neq 0$ (Abraham et al., 2008).

Despite the critical importance of statistical power for drawing reasonable inferences, numerous empirical reviews paint a sobering picture of the state of the psychological literature. Since Cohen's (1962) prescient review, a meta-scientific literature has amassed (e.g., Richard et al., 2003; Rossi, 1990), which suggests that the majority of psychological research is underpowered to find the most typically observed effect sizes (i.e., r = .20 to .30), much less even smaller (r = .10) effects. In the most recent comprehensive reviews available of the psychological and psychiatric literatures (Dumas-Mallet et al., 2017; Szucs et al., 2017), the mean power to detect effects of r = .10 and .30 was 17% and 49%, respectively.

The magnitude of concern that this trend raises for the interpretability of the psychological literature cannot be overstated. These results suggest that even when a true, medium-sized effect exists in the population, the average study is only sufficiently powered to find this effect around half of the time. The case is even more dire for small effects: the average psychology study is powered to find evidence for a small effect less than one quarter of the time even when the effect exists! To make matters worse, effect sizes associated with statistical significance in under-powered samples may be substantial over-estimates of the population effect size, as larger effect sizes are often required to achieve statistical significance with smaller

samples (e.g., Fraley et al., 2014; Kühberger et al., 2014). This also has the potential to create a harmful cycle, such that an overestimated effect size from one under-powered study may be cited as evidence in the calculation of a power analysis for subsequent studies. Moreover, the accretion of such studies can ultimately influence future research, prevention, and policy decisions.

The Case of Laboratory Aggression

Several recent reviews suggest that power is a substantial concern in the laboratory aggression literature, where the largest meta-analytic effect sizes (i.e., psychopathy and lab aggression) are in the realm of r = ~.20 (Hyatt et al., 2019). In a *p*-curve analysis (Simonsohn et al., 2014a) spanning the entirety of studies using the seminal Taylor Aggression Paradigm (TAP; Taylor, 1967)[1], the mean effect size for hypothesized main effects (e.g., effect of an individual difference variable or experimental condition) across all studies was Cohen's d = .37 (or r = .18) and d = .10 (or r = .05) for interaction effects (West et al., 2020). Of note, this average main effect size is consistent with the average effect size (r = .20) found in the recent replication effort by the Open Science Collaboration (2015) and can be characterized as small in traditional conventions (Cohen, 1969) or medium in magnitude based on more recent recommendations (Funder et al., 2019; Gignac et al., 2016). Assuming α = .05, this translates to mean power of 58% and 12% to identify main (mean N = 127) and interaction (mean N = 134) effects, respectively, in laboratory aggression work (West et al. 2020). Although this *p*-curve analysis found that statistical power to test a focal hypothesis for a main or interaction effect using lab aggression methodology has improved over the last decade (i.e., mean power 2000-2009 = 10% vs. 2010-2020 = 44%), it remains unacceptably low.

[1]Due to lack of requisite information, not every study that used the Taylor Aggression Paradigm could be included in this review. Please see West et al. (2020) for more information.

The Current Study

The laboratory aggression literature, like the rest of the psychological literature, is generally not well-powered to find main effects and *severely* underpowered to find interaction effects. The aim of the current effort was to provide a useful resource to help counter this trend. Using data simulations, we present an empirically based power primer that laboratory aggression researchers can use to guide their decisions regarding the number of subjects that they should plan to collect based on study design and relations among predictors. Although free statistical power calculators exist (e.g., G*Power; Faul et al., 2009; *pwr* in R; Champeley, 2020), we believe the current work has many advantages for laboratory aggression researchers for several reasons.

First, we use effect size estimates found to be most plausible in recent comprehensive reviews (Hyatt et al., 2019; West et al., 2020). In turn, we can make sample size recommendations based on the most up-to-date research on effect sizes in laboratory aggression research. Second, while free software packages appear intuitive and user-friendly, decisions about the type of statistical test and power analysis to run are not always as straightforward as they may seem. The effect sizes required for these programs (e.g., Cohen's f^2) are not easily interpreted within the metrics most commonly observed in the literature (e.g., β regression coefficients). These programs also recommend arbitrary values for what constitute small, medium, or large effects (Correll et al., 2020). Additionally, measurement error is rarely taken into account, which becomes increasingly problematic when testing for interactions which necessarily contain more measurement error than main effects (e.g., reliability of an interaction term created from two continuous variables, each with an $\alpha = .70$, has an $\alpha = .49$). Thus, we provide guidelines for desired sample sizes based on the type of research question being asked

(e.g., does trait X correlate with laboratory aggression? Are there differences in aggression by experimental condition? Does trait X interact with condition Y to engender particularly elevated levels of aggression?) and the nature of the variables of interest (e.g., continuous, dichotomous).

Based on the most commonly observed designs across the history of the laboratory aggression literature (e.g., West et al., 2020), we present power analyses for 1) main effects of a continuous predictor (e.g., personality trait), 2) main effects of experimental condition, 3) two-way interactions between predictors of various interrelations, and 4) three-way interactions between a continuous predictor and two dichotomous predictors (e.g., experimental condition, gender) of various interrelations. We present findings for both one- and two-tailed tests, given that both types of tests may be sensible given a researcher's hypothesis, but present the results of the two-tailed tests as primary given that this type of test is predominant in the psychological literature.

Methods

Effect Size Estimates

Main effects. The effect size estimates for the simulation analyses were derived from several reviews and meta-analyses in the interest of the selecting effects representative of those likely to be found in future studies. The effect size estimate for the main effect of a continuous predictor on lab aggression was taken from the meta-analysis by Hyatt and colleagues (2019) on personality traits and laboratory aggression. The largest effect identified was r = .23 between psychopathy – lab aggression, and effects of similar magnitudes were observed for narcissism (r = .20), Five Factor Model Agreeableness (r = -.20), and sadism (r = .19). The effect size estimate for the main effect of experimental condition was taken from the *p*-curve analysis by West and colleagues (2020), where the mean main effect of experimental manipulation was Cohen's d =

.27. Importantly, this does not represent the raw mean effect size observed across these studies, but rather the effect size that best fits the observed data after accounting for publication bias given the estimates of the *p*-curve (see Simonsohn et al., 2014b). Based on these observed effects, we elected to provide power estimates for main effects of r = .10 (d = .20), r = .15 (d = .30), r = .20 (d = .41), and r = .30 (d = .63) as reasonable estimates of the range of main effect sizes likely to be observed in this literature.

Interaction effects. The simulation analyses for the two- and three-way interaction effects required effect size estimates for the interaction effects themselves, effect size estimates for the main effects of the predictors, as well as correlations among the two or three predictors. The effect size estimates for the interactions were estimated using the publicly available coding scripts (https://osf.io/h9c85/) generated for the *p*-curve analysis by West and colleagues (2020); the mean effects for the two- and three-way interactions were r = .06 (d = .12) and r = .01 (d = .02), respectively. Thus, for two-way interactions, we elected to provide estimates for r = .05 (d = .10), r = .10 (d = .20), r = .20, (d = .41). In selecting these two-way interaction effect size estimates, we recognize that these larger values are over-estimations of the effects typically observed in this literature.

For three-way interactions, we provide estimates for r = .01 (d = .02), r = .05 (d = .10), and r = .10 (d = .20). In selecting these three-way interaction effect size estimates, we recognize that this lower bound is relatively unlikely to be applicable, given that researchers tend to be uninterested in such small effects, and the upper bound is not a reasonable *a priori* effect size estimate to anticipate given that three-way interaction effects of those magnitudes are very rarely observed.

For the main effect size estimates in the interaction analyses, $r = .20$ was used for the main effect of a continuous predictor based on the meta-analysis by Hyatt and colleagues (2019). Similarly, based on the results of the work by West and colleagues (2020), $r = .10$ was used for the main effect of condition as the best estimate of the effect of condition after correcting for publication bias. Finally, in a large, internal dataset (N = 1,790; Lasko et al., 2021), the gender difference in TAP scores was $d = .23$ ($r = .11$) such that males exhibited higher aggression scores than females. Thus, $r = .10$ appears a reasonable effect size estimate for both types of dichotomous predictors of interest, and results are generalizable to both continuous variable*condition and continuous variable*gender interactions.

We assumed random assignment to experimental condition and thus the relation between condition and other variables of interest (i.e., continuous predictor, gender) would be $r = 0$. Given that there is likely to be variation in the magnitude of effect size between the other predictors (i.e., two continuous predictors or continuous predictor – gender) depending on the variables of interest, several values were used ranging from $r = 0$ to .50, including $r = 0$, $r = .10$ ($d = .20$), $r = .30$ ($d = .63$), and $r = .50$ ($d = 1.16$). We believe this covers the range of correlations between predictors likely to be observed in this literature, ranging from no relation ($r = 0$) to a very large relation ($r = .50$) that may be observed between conceptually and empirically similar continuous predictors, like subcomponents of a multidimensional construct (e.g., antagonism and disinhibition components of psychopathy). Notably, this range of values captures those found for the largest gender differences in psychological variables identified to date (Hyde, 2014).

Simulation Analyses

Simulations were conducted using the R statistical analysis software (Version 4.0.2; R Core Team, 2020) in RStudio (RStudio Team, 2020). The nature of these analyses is such that there is no dataset to make available, but all code to reproduce our simulation analyses is available at the following link (https://osf.io/vcq3s/). Separate simulations were run for each effect of interest (i.e., main effects, two-way interactions, and three-way interactions at a range of effect sizes), and each simulation was broken into two parts. In part 1, a null effect was simulated to identify an effect cutoff consistent with the nominal Type 1 error rate of 5%. Hypothetical predictor variables with known distributions and associations (r_{ij}) were generated with given sample sizes. Continuous variables were normally distributed ($Y = 0$, $SD = 1$) while dichotomous variables were generated with a 50% probability of generating either value of 0 or 1. Measurement error was then introduced to the predictor variables based on a given measurement reliability. For continuous variables, measurement reliability (α) was set to .85, consistent with the upper end of meta-analytic estimates for trait measures (e.g., Miller et al., 2018; Peterson, 1994). This is likely a generous estimate and thus provides a power estimate that assumes excellent measurement precision. Standard error of measurement ($SD\sqrt{1-\alpha}$) was calculated and used to generate error around the continuous variable. The dichotomous predictor variable was always simulated without measurement error as it was intended to represent assigned experimental condition.

A continuous, normally distributed outcome variable (Y) was then generated with a known association with the predictor variables ($Y = \beta_1 X_1 + \ldots + \beta_j X_j$) consistent with the null hypothesis. When evaluating power for interaction effects, only main effects were included at this stage (i.e., no interaction effects). The residual variance of the outcome ($1 - R^2$) was

calculated as a function of the β coefficients and the association between predictors (Cohen et al., 2002):

$$R^2 = \sum(\beta_i^2) + 2\sum \beta_i \beta_j r_{ij} \quad (1)$$

The standard deviation of the outcome variable was then identified using: SD = $\sqrt{(1 - R^2)}$. Measurement error was also added to the outcome using the same measurement reliability parameter (i.e., α = .85). For all simulations 2,000 of such datasets were generated and a regression based on the alternative hypothesis was run. Consistent with the selected Type 1 error rate of .05, a statistically significant effect parameter was observed roughly 5% of the time for all simulations. From this, a cutoff representing the 97.5th percentile for two-tailed simulations (or the 95th percentile for one-tailed simulations) of the 2,000 β coefficients of interest was saved. This cutoff represents the value above which only 2.5% (or 5%) of the coefficients would fall given a true null effect. This cutoff was used in part 2 of the simulation to hold the Type 1 error rate constant at 5% while power was being evaluated.

In part 2 of the simulation, 2,000 new datasets were generated. All variables were generated in the same manner, with the key exception being that the tested effect of interest was added to the model when generating the outcome variable. The outcome was again regressed on the predictors in each of the generated datasets and power was determined by calculating the percentage of the 2,000 β coefficients of interest that fell above the 97.5% (or 95%) cutoff identified in the Type 1 error simulations.

Realistic parameters were chosen for all untested effects within the higher complexity (i.e., two-way and three-way interactions) simulations. In the two-way interaction simulations, the main effect coefficients were assigned as β = .1 for the dichotomous predictor variable and β = .2 for the continuous variable. All variables were standardized so these could be interpreted as

partial correlation coefficients. The three-way interactions were generated with two dichotomous variables and one continuous variable, and again the dichotomous variables had main effects of β = .1 while the continuous predictor was assigned a main effect of β = .2. The two-way interaction effects included as part of the three-way interaction simulations were given rationally identified magnitudes. The interaction between the dichotomous predictors was assigned as null while the interactions between the dichotomous predictors and the continuous predictors had magnitudes of β = .05 and β = .1.

Results

Main Effects

Results of the simulation-based power analyses for main effects can be found in Table 1. As expected, power estimates increase with sample size, and one-tailed directional tests have more power than two-tailed tests. An important pattern to note is the variability in power across the various effect size estimates. For example, if researchers wish to locate a main effect of a continuous variable (r = ~.20) of a similar magnitude to the largest effects found (i.e., psychopathy – lab aggression) in the meta-analysis by Hyatt and colleagues (2019) using a two-tailed test, they should plan to collect at least N = 200 to achieve ~80% power. Researchers interested in other constructs with less established links to aggression *should not* assume an effect size of this magnitude. Researchers who expect to locate a main effect of a continuous variable consistent with the smaller effect sizes (e.g., r = ~.10; impulsivity – lab aggression) found in the meta-analysis by Hyatt and colleagues (2019) using a two-tailed test should plan to collect at least N = 1,000 to achieve ~80% power.

Researchers who expect to locate a main effect of condition (d = ~.30) of similar magnitude found in the meta-analysis by West and colleagues (2020) using a two-tailed test

should plan to recruit at least N = 400 to achieve ~80% power. Researchers hypothesizing a smaller, yet still empirically common effect size (*d* = ~.20; Gignac et al., 2016) should plan to collect at least N = 1,000 to achieve ~80% power.

Two-Way Interactions

Results of the simulation-based power analyses for two-way interaction effects can be found in Table 2. As before, power estimates increase with sample size and one-tailed tests yield more power than two-tailed tests. Additionally, a pattern was observed such that across effect size estimates and most sample sizes, statistical power decreases somewhat as the correlations between the predictor variables increased.

Researchers who expect to locate an interaction effect of a magnitude typically found in this literature (i.e., *r* = ~.05; West et al., 2020) with a two-tailed test should plan to collect at least N = 5,000 to achieve ~80% power across all tested predictor variable correlations. Even if researchers assume a two-way interaction effect size of a magnitude of double the size typically found, they should plan to collect at least N = 1,000 to achieve ~80% power across all tested predictor variable correlations.

Three-Way Interactions

Results of the simulation-based power analyses for three-way interaction effects can be found in Table 3. As before, power estimates increase with sample size and one-tailed tests yield more power than two-tailed tests. Similarly, the same pattern of power across predictor variable correlations was observed such that across effect size estimates and most sample sizes, statistical power decreased somewhat as the correlations between the predictor variables increased. Researchers who expect to locate an interaction effect of a magnitude typically found in this literature (i.e., *r* = ~.01; West et al., 2020) using a two-tailed test should plan to collect at least N

= 100,000 to achieve ~80% power across all tested predictor variable correlations. Even if researchers assume a three-way interaction effect size of a magnitude *five times larger* than the size typically observed, they should plan to collect at least N = 5,000 to achieve ~80% power.

Supplemental Materials

To present comprehensive and widely applicable tables, we ran additional simulations where we varied several of the key parameters to examine the impact on statistical power estimates. Specifically, we ran additional simulations for main effects, two-way interactions, and three way interactions using a *p*-value of .005 (see Benjamini et al., 2018; Supplemental Tables 1-3), assuming a α =.65 (Supplemental Tables 4-6) and α =.45 (Supplemental Tables 7-9) for the continuous predictors, and assuming α =.65 (Supplemental Tables 10-12) and α = 1 for the dependent variable (Supplemental Tables 13-15). Results of these simulations suggest that as expected, using a more stringent *p*-value results in substantial reductions in power estimates, especially at lower sample sizes. Additionally, simulations using lower reliability estimates for the predictors and the dependent variable tended to reduce power in the range of 1%-8%.

Discussion

In this laboratory aggression power primer, we used simulation-based power analyses to provide concrete sample size recommendations to guide study design for researchers to reliably locate main and interaction effects of various magnitudes. Importantly, this manuscript is designed to be a resource to assist in study design. It is not intended to be used such that researchers can justify their sample size *post hoc* based on the observed effect sizes, as this *post hoc* approach to power has fundamental flaws (see Gelman, 2018; Hoenig et al., 2001). Researchers interested in reflecting on the effect that a study could detect when the sample size, desired power, and alpha are fixed may consider a sensitivity power analysis (see Lakens, 2021).

We are hopeful that aggression researchers will find this resource useful for *a priori* study planning, as there is a stark contrast between these findings and sample sizes observed in the laboratory aggression literature. For example, the current analyses demonstrate that a sample of N = 200 is needed to locate a main effect of a continuous variable with a two-tailed test of the magnitude (r = ~.20) found in the meta-analysis by Hyatt and colleagues (2019) at ~80% power. Only 6 of the 123 effect sizes (4.9%) reported in this meta-analysis were derived from samples of this size or larger. Similarly, the current results show that a sample of N = 400 is needed to locate a main effect of condition with a one-tailed test at the magnitude (d = ~.30) reported in the meta-analysis by West and colleagues (2020) at ~80% power, but only 1 of the 41 (2.4%) effect sizes located in this manuscript were derived from samples of this threshold.

The case is much worse for interaction effects. The current analyses demonstrate that a sample of N = 5,000 is needed to locate a two-way interaction effect with a two-tailed test of the magnitude (r = ~.05) reported in the meta-analysis by West and colleagues (2020) at >80% power. None of the 55 reviewed studies reported sample sizes that met this benchmark, which is not surprising given the time and resource intensive nature of laboratory aggression work. Similarly, if one expects a three-way interaction of the magnitude (r = .01) derived from the data and coding scripts provided by West and colleagues (2020), a sample of N = 100,000 is needed to achieve >80% power with a two-tailed test, a benchmark that none of the reviewed studies approach. The practical upshot of this finding is that we urge lab aggression researchers to consider an indefinite moratorium on testing three-way interactions and only propose them in cases where a massive interactive effect can be expected. Based on the pattern of effect sizes observed in this literature to date, rationale for such an effect size must be unusually strong to justify such a data collection effort.

Ultimately, the current results underscore that the vast majority of lab aggression work is underpowered, and the case is especially dire for interaction effects (e.g., narcissism by ego threat condition in the prediction of aggression). This also suggests that many aggression studies were doomed to the file drawer from their outset, as they were too underpowered to detect their intended effect – representing a considerable waste of time and resources. Nonetheless, there are many significant interaction effects published in the literature, which suggests the presence of Type I errors. Beyond significance testing, these underpowered tests in the published literature are also problematic in that they may yield overestimates of the true effect found in subsequent, higher-powered samples (e.g., Ioannidis, 2008; Kühberger et al., 2014; Schönbrodt et al., 2013).

Implications for Lab Aggression Researchers

We hope the current work can guide decision-making around laboratory aggression research and offer several suggestions for moving forward. First, researchers should take seriously the issue of whether a given study *should* be run given scholarly potential and practical limitations to data collection. There are important costs to conducting underpowered analyses (Lilienfeld et al., 2020), including the risk of Type I/II errors that may hold implications for decisions made by preventionists and interventionists. There are also significant human costs to participants and researchers who devote time and energy to a data collection effort that may be, at best, not meaningfully informative, and at worst, consequentially harmful to applied efforts. Second, if the study is deemed worth running, it is worth considering if laboratory aggression approaches represent the best methodological option. There are many operationalizations of aggression available, including self-report measures of trait aggressiveness (Buss & Perry, 1992; Chester & West, 2020), self-report of behavioral instances of aggression (Straus, Hamby, Boney-McCoy, & Sugarman, 1996), ecological momentary assessment (Murray et al., 2020), informant

report (Tackett & Ostrov, 2010), hypothetical responding to vignettes (Crick & Dodge, 1996), behavioral observation (Bandura, Ross, & Ross, 1961), and structured patient chart review (Brown, Goodwin, Ballenger, Goyer, & Major, 1979). By using multiple aggression methodologies, researchers may be able to study aggression in a manner that is ultimately more holistic, multi-faceted, pragmatic, and replicable. Third, if researchers determine that laboratory aggression methodology is the best fit for the research question at hand due to its unique strengths (i.e., internal validity via control over confounds), we emphasize the importance of also following other forms of best practice, including pre-registration of analytic strategy (Elson et al., 2014; Hyatt et al., 2019) and hypotheses, and use of validated experimental manipulations and standardized protocols (e.g., Chester et al., 2019a, 2019b). For scholars that want to leverage the unique benefits of lab aggression measures and test small main effects or interactions, it may be advisable to form multi-site, multi-investigator collaborations that reduce the experimenter burden and decrease the time it takes to collect a sufficiently powered sample (e.g., Many Labs; Klein et al., 2018). We also refer researchers to novel quantitative work on alternative measures of effect size in moderation analyses that may be useful in providing multiple perspectives on the amount of variance explained by an interaction term (Liu et al., 2020).

The current results also highlight the discrepancies in power between one- and two-tailed tests, such that one-tailed tests are associated with higher statistical power. The issue of which type of test is most appropriate is a long-standing controversy in psychological science (e.g., Eysenck, 1960) and we refrain from weighing in on these larger statistical and philosophical arguments here. Several sets of criteria for determining when a one-tailed test is appropriate have been proposed (e.g., Kimmel, 1957; Ruxton et al., 2010); these are important to consider given the clear costs to one-tailed tests that may make them inadvisable (e.g., inability to draw

inferences about effects in the non-hypothesized direction), especially in exploratory work. In many cases, however, researchers using laboratory aggression paradigms have clear, *a priori* hypotheses (e.g. alcohol ingestion will be related to greater aggression), and thus one-tailed tests may be appropriate. By pre-registering directional hypotheses, analytic approaches, multiple comparison controls, etc., researchers can appropriately use one- rather than two-tailed tests, since pre-registrations give confidence to readers that such a decision was not biased *post hoc* by the observed results.

Lastly, we direct interested readers to a newly developed R package called InteractionPoweR (Baranger et al., 2021) and the associated shiny app (Finsaas et al., 2021). This excellent set of resources also uses simulation-based power analyses to provide power estimates for two-way interactions. While we attempted to be relatively exhaustive, we encourage use of these resources, especially if other combinations of parameters are better suited for a particular research question (e.g., extremely highly correlated predictors; skewed variable distributions). However, we strongly advise that researchers make use of the effect size estimates provided herein or to justify why a larger effect was anticipated, especially for tests of interaction where extremely small effects are typical.

Broad Implications

Although we have discussed these results in the context of laboratory aggression research, there are clear implications of this work beyond this literature. Behavioral tasks are used (and prized) broadly in the social sciences, and we argue that relatively small main effects and very small interaction effects found in laboratory aggression meta-science work are likely the rule for this type of methodology, not the exception. Given that effect sizes of similar magnitudes are common across psychology (e.g., Aguinis et al., 2005; O'Boyle et al., 2019;

Plonsky et al., 2014), psychological scientists across content areas could read this manuscript substituting in their topic of choice and the implications/recommendations would be unchanged. Moreover, underpowered work can contribute to harmful cycles that extend beyond the scientific domain: underpowered work gets included in reviews and meta-analyses, these meta-scientific compilations inform policy, policy informs prevention/intervention efforts, and these efforts impact societal well-being (Banks et al., 2012).

In closing, we recognize that the sample size recommendations made herein may be a difficult pill to swallow for laboratory aggression researchers (and beyond), especially given the contrast between these recommendations and typical sample sizes employed. In the interest of transparency, we also note that members of our research group have tested interactions in samples that, in light of the current findings, were underpowered to locate such an effect reliably (e.g., Hyatt et al., 2017; Vize et al., 2021). Nonetheless, we reiterate long-standing (Cohen, 1962) calls for attention to this crucial element of study planning and hope that laboratory aggression researchers will lead the charge in terms of solidifying the credibility of our methods. Given the importance of aggression as a public health concern, we believe the benefits of reliable and well-powered research on the causes and consequences of this behavior far outweigh the costs (LeBel, Campbell, & Loving, 2017).

References

Abraham, W. T., & Russell, D. W. (2008). Statistical power analysis in psychological research. *Social and Personality Psychology Compass*, *2*, 283-301.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: a 30-year review. *Journal of Applied Psychology*, *90*, 94-107.

Bandura, A., Ross, D., & Ross, S. A. (1961). Transmission of aggression through imitation of aggressive models. *The Journal of Abnormal and Social Psychology*, *63*, 575-582.

Baranger, D.A.A., Finsaas, M.C., Goldstein, B.L., Vize, C. E., Lynam D. R., & Olino, T.M. InteractionPoweR: Power analyses for interaction effects in cross-sectional regressions. (2021) https://github.com/dbaranger/InteractionPoweR.

Baron, R. A. (1977). *Human aggression*. New York: Plenum.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Cesarini, D. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6-10.

Brown, G. L., Goodwin, F. K., Ballenger, J. C., Goyer, P. F., & Major, L. F. (1979). Aggression in humans correlates with cerebrospinal fluid amine metabolites. *Psychiatry Research*, *1*, 131-139.

Buss, A. H., & Perry, M. (1992). The aggression questionnaire. *Journal of Personality and Social Psychology*, *63*, 452-458.

Champely, S. (2020). pwr: Basic functions for power analysis. R package version 1.2-2. https://CRAN.R-project.org/package=pwr

Chester, D. S., & Lasko, E. (2019a). Construct validation of experimental manipulations in social psychology: Current practices and recommendations for the future. Manuscript under review, available at: https://psyarxiv.com/t7ev9

Chester, D. S., & Lasko, E. N. (2019b). Validating a standardized approach to the Taylor Aggression Paradigm. *Social Psychological and Personality Science*, *10*, 620-631.

Chester, D. S. & West, S. J. (2020). Trait aggression is primarily a facet of antagonism: Evidence from dominance, latent correlational, and item-level analyses. *Journal of Research in Personality*, manuscript in press.

Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2020). Avoid Cohen's 'Small', 'Medium', and 'Large' for Power Analysis. *Trends in Cognitive Sciences*, *24*, 200–207. https://doi.org/10.1016/j.tics.2019.12.009

Crick, N. R., & Dodge, K. A. (1996). Social information-processing mechanisms in reactive and proactive aggression. *Child Development*, *67*, 993-1002.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *The Journal of Abnormal and Social Psychology*, *65*, 145-153.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. San Diego, CA: Academic Press

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F., & Munafò, M. R. (2017). Low statistical power in biomedical science: a review of three human research domains. *Royal Society Open Science*, *4*, 160254.

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*, 419-432.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160.

Finsaas, M.C., Baranger, D.A.A., Goldstein, B.L., Vize, C. E., Lynam D. R., & Olino T.M. (2021). InteractionPoweR Shiny App: Power Analysis for Interactions in Linear Regression. Retrieved 9/1/2021.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS one*, *9*, e109019.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*, 156-168.

Gelman, A. (2018, September 24). Don't calculate post-hoc power using observed estimate of effect size. Statistical Modeling, Causal Inference, and Social Science. https://statmodeling.stat.columbia.edu/2018/09/24/dont-calculate-post-hoc-power-using-observed-estimate-effect-size/

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74-78.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1-20.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician, 55*, 19-24.

Hyatt, C. S., Berke, D. S., Miller, J. D., & Zeichner, A. (2017). Do beliefs about gender roles moderate the relationship between exposure to misogynistic song lyrics and men's female-directed aggression?. *Aggressive Behavior*, *43*, 123-132.

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.

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

Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology*, *65*, 373-398.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, 640-648.

Kimmel, H. D. (1957). Three criteria for the use of one-tailed tests. *Psychological Bulletin*, *54*, 351-353.

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLOS ONE*, *9*, e105825.

Lakens, D. (2021, January 4). Sample Size Justification. https://doi.org/10.31234/osf.io/9d3yf

Lasko, E.N., & Chester, D.S. (2021). Measurement invariance and item response theory analysis of the Taylor Aggression Paradigm. *Assessment*, manuscript in press.

LeBel, E. P., Campbell, L., & Loving, T. J. (2017). Benefits of open and high-powered research outweigh costs. *Journal of Personality and Social Psychology*, *113*, 230-243.

Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology/Psychologie canadienne*.

Liu, H., & Yuan, K. H. (2020). New measures of effect size in moderation analysis. *Psychological Methods*.

McCarthy, R. J., & Elson, M. (2018). A conceptual review of lab-based aggression paradigms. *Collabra: Psychology*, *4*.

Miller, B. K., Nicols, K. M., Clark, S., Daniels, A., & Grant, W. (2018). Meta-analysis of coefficient alpha for scores on the Narcissistic Personality Inventory. *PloS one*, *13*, e0208331.

Murray, A. L., Eisner, M., Ribeaud, D., & Booth, T. (2020). Validation of a brief measure of aggression for experience sampling research: The Aggression-ES-A. Manuscript under review, available at https://psyarxiv.com/un6hv/

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.

O'Boyle, E., Banks, G. C., Carter, K., Walter, S., & Yuan, Z. (2019). A 20-year review of outcome reporting bias in moderated multiple regression. *Journal of Business and Psychology*, *34*, 19-37.

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

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of consumer Research*, *21*, 381-391.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878-912.

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.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years?. *Journal of Consulting and Clinical Psychology, 58*, 646-656.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. http://www.rstudio.com/.

Ruxton, G. D., & Neuhäuser, M. (2010). When should we use one-tailed hypothesis testing?. *Methods in Ecology and Evolution, 1*, 114-117.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize?. *Journal of Research in Personality, 47*, 609-612.

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.

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.

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

Straus, M. A., Hamby, S. L., Boney-McCoy, S., & Sugarman, D. B. (1996). The revised conflict tactics scales (CTS2) development and preliminary psychometric data. *Journal of Family Issues, 17*, 283-316.

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.

Tackett, J. L., & Ostrov, J. M. (2010). Measuring relational aggression in middle childhood in a multi-informant multi-method study. *Journal of Psychopathology and Behavioral Assessment, 32,* 490-500.

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.

Tedeschi, J. T., & Quigley, B. M. (1996). Limitations of laboratory paradigms for studying aggression. *Aggression and Violent Behavior, 1,* 163-177.

Vize, C. E., Miller, J. D., Collison, K. L., & Lynam, D. R. (2021). Untangling the relation between narcissistic traits and behavioral aggression following provocation using an FFM framework. *Journal of Personality Disorders, 35,* 299-318.

Warburton, W. A., & Bushman, B. J. (2019). The competitive reaction time task: The development and scientific utility of a flexible laboratory aggression paradigm. *Aggressive Behavior*.

West, S. J., Hyatt, C. S., Miller, J. D., & Chester, D. S. (2020). *p*-Curve analysis of the Taylor Aggression Paradigm: Estimating evidentiary value and statistical power across 50 years of research. *Aggressive Behavior*, manuscript in press.

Zeichner, A., Frey, F. C., Parrott, D. J., & Butryn, M. F. (1999). Measurement of laboratory aggression: A new response-choice paradigm. *Psychological Reports, 85,* 1229-1237.

Table 1

Power estimates for main effects of a continuous predictor and experimental condition

Continuous Predictor								
	Two-tailed $p < .05$				One-tailed $p < .05$			
N	*r* = .10	*r* = .15	*r* = .20	*r* = .30	*r* = .10	*r* = .15	*r* = .20	*r* = .30
50	9%	15%	26%	51%	18%	26%	39%	64%
100	17%	25%	46%	82%	24%	40%	60%	90%
150	20%	39%	63%	93%	30%	52%	74%	96%
200	25%	52%	79%	98%	37%	64%	86%	100%
250	33%	60%	86%	99%	43%	73%	92%	100%
300	39%	69%	91%	100%	50%	78%	95%	100%
350	38%	76%	93%	100%	53%	85%	97%	100%
400	47%	82%	96%	100%	61%	89%	98%	100%
450	52%	85%	98%	100%	64%	92%	99%	100%
500	57%	87%	99%	100%	68%	92%	99%	100%
750	72%	96%	100%	100%	81%	98%	100%	100%
1,000	83%	100%	100%	100%	90%	100%	100%	100%
2,500	100%	100%	100%	100%	100%	100%	100%	100%
Experimental Condition								
	Two-tailed $p < .05$				One-tailed $p < .05$			
N	*d* = .20	*d* = .30	*d* = .41	*d* = .63	*d* = .20	*d* = .30	*d* = .41	*d* = .63
50	10%	19%	24%	53%	16%	26%	39%	65%
100	16%	30%	46%	85%	25%	42%	59%	90%
150	23%	45%	63%	96%	32%	56%	77%	98%
200	25%	48%	75%	98%	36%	62%	84%	99%
250	30%	62%	87%	100%	41%	73%	92%	100%
300	33%	71%	89%	100%	50%	82%	95%	100%
350	45%	73%	95%	100%	57%	83%	98%	100%
400	46%	78%	96%	100%	56%	86%	98%	100%
450	51%	88%	98%	100%	63%	93%	99%	100%
500	55%	89%	99%	100%	66%	93%	100%	100%
750	74%	97%	100%	100%	81%	99%	100%	100%
1,000	83%	100%	100%	100%	90%	100%	100%	100%
2,500	100%	100%	100%	100%	100%	100%	100%	100%

Note: *r* and *d* represent the main effect size for continuous predictor and experimental condition, respectively; for the continuous predictor and the dependent variable, we assumed $\alpha = .85$; for the experimental condition, we assumed $\alpha = 1$.

Table 2

Power estimates for two-way interaction effect between predictors of various interrelations

	Two-tailed $p < .05$											
N	*r* = .05, *d* = .10				*r* = .10, *d* = .20				*r* = .20, *d* = .41			
r^{ab}	0	.10	.30	.50	0	.10	.30	.50	0	.10	.30	.50
50	6%	4%	5%	5%	9%	9%	8%	7%	20%	25%	24%	17%
100	5%	6%	6%	6%	16%	14%	14%	14%	45%	47%	48%	43%
150	8%	9%	7%	8%	21%	20%	22%	17%	60%	62%	63%	54%
200	10%	12%	10%	9%	30%	26%	24%	24%	79%	78%	76%	71%
250	11%	10%	10%	11%	33%	31%	31%	26%	87%	85%	84%	82%
300	12%	14%	14%	12%	40%	37%	35%	31%	92%	93%	91%	86%
350	14%	13%	12%	13%	41%	40%	38%	35%	94%	95%	94%	92%
400	16%	14%	16%	14%	49%	50%	46%	42%	97%	98%	96%	93%
450	17%	20%	17%	15%	49%	52%	55%	43%	98%	98%	98%	96%
500	19%	19%	21%	16%	56%	62%	54%	52%	99%	100%	99%	98%
750	24%	28%	24%	24%	73%	75%	69%	65%	100%	100%	100%	100%
1,000	30%	32%	31%	28%	86%	84%	86%	78%	100%	100%	100%	100%
2,500	69%	66%	65%	61%	100%	100%	100%	100%	100%	100%	100%	100%
5,000	92%	92%	92%	87%	100%	100%	100%	100%	100%	100%	100%	100%
7,500	98%	99%	98%	95%	100%	100%	100%	100%	100%	100%	100%	100%
10,000	100%	100%	100%	99%	100%	100%	100%	100%	100%	100%	100%	100%
	One-tailed $p < .05$											
N	*r* = .05, *d* = .10				*r* = .10, *d* = .20				*r* = .20, *d* = .41			
r^{ab}	0	.10	.30	.50	0	.10	.30	.50	0	.10	.30	.50
50	10%	9%	9%	9%	16%	17%	14%	13%	31%	37%	35%	31%
100	13%	12%	11%	11%	25%	24%	22%	21%	58%	59%	61%	55%
150	15%	13%	16%	13%	31%	31%	30%	25%	71%	74%	73%	69%
200	18%	16%	16%	15%	39%	39%	35%	32%	86%	87%	86%	81%

250	20%	18%	17%	17%	46%	45%	41%	37%	92%	91%	92%	89%
300	19%	22%	20%	17%	49%	49%	47%	41%	96%	96%	95%	92%
350	24%	20%	21%	19%	53%	57%	54%	47%	97%	98%	96%	96%
400	25%	26%	24%	20%	61%	62%	56%	53%	98%	99%	99%	97%
450	25%	28%	26%	23%	60%	65%	66%	53%	99%	99%	99%	98%
500	28%	26%	28%	23%	67%	72%	64%	62%	100%	100%	100%	99%
750	35%	38%	34%	32%	84%	83%	80%	76%	100%	100%	100%	100%
1,000	40%	46%	42%	40%	92%	91%	91%	86%	100%	100%	100%	100%
2,500	78%	77%	74%	71%	100%	100%	100%	100%	100%	100%	100%	100%
5,000	96%	96%	95%	92%	100%	100%	100%	100%	100%	100%	100%	100%
7,500	99%	100%	99%	98%	100%	100%	100%	100%	100%	100%	100%	100%
10,000	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Note: *r* and *d* values represent the effect size of the interaction term and these can be interpreted as partial *r*s; r^{ab} represents the correlation between the two predictor variables; for the continuous predictors and the dependent variable, we assumed $\alpha = .85$; in this table, we present power estimates for analyses with two continuous predictors (both $\alpha = .85$), given that these results were very highly overlapping with analyses with one continuous predictor ($\alpha = .85$) and one dichotomous predictor ($\alpha = 1$).

Table 3

Power estimates for three-way interaction effects between two dichotomous predictors and a continuous predictor of various interrelations

	Two-tailed $p < .05$											
N	*r* = .01, *d* = .02				*r* = .05, *d* = .10				*r* = .10, *d* = .20			
r^{ab}	0	.10	.30	.50	0	.10	.30	.50	0	.10	.30	.50
50	3%	3%	3%	3%	5%	5%	5%	4%	9%	8%	7	7
100	3%	3%	3%	3%	6%	6%	6%	6%	13%	13%	14%	12%
150	3%	3%	3%	3%	8%	7%	8%	7%	21%	19%	19%	17%
200	3%	3%	3%	3%	9%	9%	9%	9%	27%	28%	22%	18%
250	3%	4%	4%	3%	12%	11%	11%	11%	34%	32%	28%	26%
300	4%	3%	4%	3%	11%	14%	12%	8%	39%	39%	31%	29%
350	3%	4%	4%	4%	14%	12%	16%	13%	44%	41%	42%	34%
400	4%	4%	4%	3%	16%	14%	14%	14%	52%	49%	50%	36%
450	4%	3%	4%	4%	17%	17%	15%	14%	55%	55%	52%	43%
500	4%	4%	4%	4%	21%	19%	17%	15%	57%	58%	56%	48%
750	4%	5%	4%	4%	28%	26%	21%	21%	76%	76%	72%	68%
1,000	5%	5%	5%	4%	35%	35%	27%	25%	86%	85%	86%	77%
2,500	6%	7%	7%	7%	68%	71%	64%	51%	100%	100%	100%	99%
5,000	11%	10%	9%	10%	92%	93%	92%	86%	100%	100%	100%	100%
7,500	16%	14%	11%	11%	99%	98%	99%	97%	100%	100%	100%	100%
10,000	13%	16%	16%	14%	100%	100%	100%	99%	100%	100%	100%	100%
100,000	88%	88%	83%	78%	100%	100%	100%	100%	100%	100%	100%	100%
	One-tailed $p < .05$											
N	*r* = .01, *d* = .02				*r* = .05, *d* = .10				*r* = .10, *d* = .20			
r^{ab}	0	.10	.30	.50	0	.10	.30	.50	0	.10	.30	.50
50	6%	6%	6%	6%	9%	9%	9%	8%	15%	14%	11%	12%
100	6%	6%	6%	6%	12%	14%	12%	10%	23%	22%	23%	20%
150	6%	6%	6%	6%	14%	14%	14%	12%	29%	30%	29%	26%

200	7%	7%	7%	6%	18%	16%	17%	15%	39%	38%	36%	31%
250	7%	7%	7%	7%	18%	20%	17%	17%	46%	45%	39%	38%
300	7%	7%	6%	6%	20%	20%	18%	15%	48%	50%	47%	42%
350	7%	7%	7%	7%	22%	20%	22%	19%	56%	54%	53%	48%
400	7%	7%	8%	7%	25%	23%	22%	21%	63%	59%	59%	52%
450	7%	7%	7%	7%	25%	26%	22%	23%	67%	67%	63%	56%
500	7%	8%	7%	8%	30%	29%	29%	22%	70%	70%	67%	63%
750	9%	8%	9%	8%	36%	36%	34%	29%	86%	84%	80%	77%
1,000	9%	9%	9%	8%	46%	48%	40%	34%	93%	91%	91%	85%
2,500	12%	12%	11%	11%	76%	78%	74%	67%	100%	100%	100%	100%
5,000	17%	18%	17%	16%	97%	97%	96%	92%	100%	100%	100%	100%
7,500	24%	22%	19%	17%	100%	99%	99%	98%	100%	100%	100%	100%
10,000	24%	23%	22%	22%	100%	100%	100%	100%	100%	100%	100%	100%
100,000	93%	94%	90%	86%	100%	100%	100%	100%	100%	100%	100%	100%

Note: *r* and *d* values represent the effect size of the interaction term; r^{ab} represents the correlation between the continuous predictor and one of the dichotomous variables; we only provide estimates for a series of r^{ab} values given that r^{ac} and r^{bc} would be = 0 since three-way interactions will likely include experimental conditions which are likely unrelated to the continuous predictor or gender; for the continuous predictors and the dependent variable, we assumed $\alpha = .85$.