Dear Author,

Thank you for using BluPencil, the online Proofing System.

While this PDF is auto-generated on the fly, the changes made here are saved as comments or annotations, and figures may be of low resolution. This version will not be as-is published online nor printed. Once the corrections are registered on BluPencil, the same will be validated against the journal-specific styles by our editors, updated, and then only the final typeset PDF will be generated for publication.

Thank you,
Team BluPencil

Professional Psychology: Research and Practice

ISSN: 0735-7028
https://doi.org/10.1037/pro0000386

Providing a Lower-Bound Estimate for Psychology's "Crud Factor": The Case of Aggression

Christopher J. Ferguson[1] and Moritz Heene[2]
[1] Department of Psychology, Stetson University
[2] Department of Psychology, Munich Center of the Learning Sciences, Ludwig-Maximilians Universität München

When conducting research on large data sets, statistically significant findings having only trivial interpretive meaning may appear. Little consensus exists whether such small effects can be meaningfully interpreted. The current analysis examines the possibility that trivial effects may emerge in large datasets, but that some such effects may lack interpretive value. When such results match an investigator's hypothesis, they may be over interpreted. The current study examines this issue as related to aggression research in two large samples. Specifically, in the first study the National Longitudinal Study of Adolescent to Adult Health (AddHeath) dataset was used. Fifteen variables with little theoretical relevance to aggression were selected, then correlated with self-reported delinquency. For the second study, the Understanding Society database was used. As with Study 1, 14 nonsensical variables were correlated with conduct problems. Many variables achieved "statistical significance" and some effect sizes approached or exceeded $r = .10$, despite little theoretical relevance between the variables. It is recommended that effect sizes below $r = .10$ should not be interpreted as hypothesis supportive.

Keywords: effect sizes, crud factor, aggression

In 1991, Meehl somewhat amusingly proposed the concept "crud factor." This concept suggests almost everything correlates with almost everything else in psychology and most of these correlations are not theoretically, interpretively, practically, or clinically meaningful. This concept is reiterated, in other ways such as "ambient noise" (Lykken, 1968) and the concerns that spurious correlations in "soft psychology" can mislead scholars (Standing et al., 1991). This issue of crud factor has created a sometimes fierce debate regarding how to interpret very small effect sizes which demonstrate statistical significance in large sample studies. Two decades ago, an American Psychological Association task force (Wilkinson & Task Force on Statistical Inference, 1999) argued for the importance of effect sizes in relation to *p* values in deciding the scientific and practical value of a research finding. However, with little clear guideline for *how* to interpret effect sizes, scholars can come to very different conclusions about the meaningfulness of a given effect. Such decisions are often made in alliance with a researcher's or clinician's a priori heuristic biases (Norcross et al., 2017). The current article seeks to provide one metric by which a floor threshold for crud factor results may be established. That is to say the current study elucidates a level of effect sizes from large, high-quality samples which may be particularly prone to crud and, as such, misleading interpretation.

Crud on Screens

One illustration of this confusion can be highlighted by a fierce debate among scholars that has erupted over the potential impact of screen technology on adolescent suicide and mental health. In particular, one study attracted national news attention for linking screen use to suicide and depressive symptoms particularly among girls (Twenge et al., 2018). However, effect sizes for this link were very small, in a sample of tens of thousands of adolescents. For instance, the correlation between social media use and depressive symptoms in girls was significant [$r(37, 830) = .06$], though the correlation ($r = .01$) was non-significant for boys. This raises the question of whether an effect size as small as $r = .06$ should be considered hypothesis supportive so long as it also reaches the $p < .05$ threshold.

Indeed, the conclusions of Twenge et al., soon proved controversial. Using several similar datasets, Orben and Przybylski (2019) found that the magnitude of the effect linking screens to depression and suicide is not greater than several obvious trivial relationships such as the effect size for eating bananas or wearing eyeglasses on suicide. There is no movement to warn parents of the dangers of

Christopher J. Ferguson https://orcid.org/0000-0003-0986-7519
Christopher J. Ferguson holds a PhD in psychology from the University of Central Florida. He is a professor of psychology at Stetson University. His research interests include the impact of aggressive video games and other media, sexualized media and thin-ideal media and body image.
Moritz Heene received his doctoral degree in psychology from the University of Heidelberg, Germany. He is currently full professor at the Department of Psychology, Munich Center of the Learning Sciences, Ludwig Maximilians University Munich, Germany. His areas of professional interest include simulation studies on structural equation modeling and statistical model testing, the logic of measurement and additive conjoint measurement, and quantitative genetics.
Christopher J. Ferguson conceived of the study, ran the statistics for Study 1 and wrote the first draft of the paper. Moritz Heene ran the statistics for Study 2 and co-wrote the final draft of the manuscript.
The preregistration for Study 1 is available at: https://osf.io/e9mtb/register/564d31db8c5e4a7c9694b2be. The code for Study 2 is presented at: https://osf.io/epfd5/
Correspondence concerning this article should be addressed to Christopher J. Ferguson, Department of Psychology, Stetson University, 421 N. Woodland Blvd., DeLand, Casselberry, FL 32729, United States. Email: CJFerguson1111@aol.com

bananas or eyeglasses. If we accept that these effect sizes related to bananas and eyeglasses, though statistically significant, nonetheless lack evidentiary value, does this run the risk that scholars selectively interpret or do not interpret effect sizes depending on their a priori beliefs? Other evidence suggests that these small effect sizes are highly susceptible to methods variance (Orben et al., 2019) or that longitudinal analysis did not reveal a predictive value for screen use on adolescent mental health (Heffer et al., 2019). Thus, a reliance on potentially trivial effect sizes to support a hypothesis may lead to false positive conclusions.

Another example related to aggression also has a link to screens. Whether video game violence does or does not contribute to violence has been contested for some time. Initial estimates, suggest violent games might account for effect sizes in the range of $r = .2$–.3 with aggression related outcomes (e.g., Huesmann, 2007). Yet, as some of these earlier studies were challenged (e.g., Adachi & Willoughby, 2011) effect sizes began to decline as did, arguably, standards of evidence. For instance, in one recent meta-analysis of violent game effects on youth aggression, the authors concluded that effect sizes in the range of $r = .08$ had evidentiary value (Prescott et al., 2018). Other meta-analyses with similar effect sizes have been more cautious in interpreting them as having evidentiary value (Furuya-Kanamori & Doi, 2016). By contrast, in another area of aggression research, effect sizes in the range of $r = .11$ were considered not to have evidentiary value regarding links between aggression and empathy (Vachon et al., 2014). Again, these contrasting interpretations both within and across aggression-related research fields point to confusion among scholars about how to interpret very small effect sizes.

It is worth recalling the explanatory power of such small correlations as well. A $r = .06$, such as for social media and depression in girls, using the coefficient of determination ($r^2 \times 100$) reveals that social media overlaps exactly 0.36% with the variance in depression, even assuming the observed relationship is not merely spurious due to methodological issues. The $r = .11$ relationship for aggression and empathy suggests a 1.2% overlap in variance, again assuming this is not inflated in some way. Granted, there are few clear goals on what is considered "practically significant" (though Ferguson, 2009 suggests a cut-off of $r = .2$ or 4%), but we suggest understanding the percent of explained or overlapping variance can give us some insight into whether an observed relationship is hypothesis supportive from a practical standpoint, particularly when so many such effects are used to guide policy or make claims about important behavioral effects.

It is also important to understand how, or why scholars might be biased in favor of overinterpreting crud. We do not mean to be overly critical here, but merely are speaking in regards to the idea of scientists being human and subject to incentives of various sorts. These biases may include moral biases (save the children, protect marginalized groups, etc.), including mainly liberal political biases (Redding, 2001). Well understood issue of publication bias may pressure scholars to interpret any "statistically significant" findings as hypothesis supportive, no matter how small or methodologically compromised. This may be particularly true for scholars with tenure and promotion deadlines though, of course, hardly limited to them. Hypothesis supportive findings also likely garner more attention from news media and policy makers, adding further incentives geared toward overinterpreting weak findings.

A Theory of Crud

Being human, scholars are likely to be tempted to believe an effect size is *real* so long as it crosses $p < .05$ and fits with their hypotheses. Thus, it helps to understand how and why many effect sizes are not *real* even if they are *statistically significant*. Part of the problem is that null-hypothesis significance testing only accounts for sampling error, not numerous other sources of error. Scholars may mistakenly assume that any effect that is statistically significant represents a reasonable estimate of a population effect size. However, spurious effect sizes can develop under multiple conditions and these will become statistically significant in large samples.

Orben and Lakens (2020) provide an important and insightful discussion of "crud." Of particular importance is understanding that there is disagreement in how it is defined and best conceptualized, nor is there clarity on how to detect it. Without such clarity, it is possible both for researchers to interpret any effect size as meaningful or, conversely, disregard effects they do not like as crud. As the use of large datasets becomes increasingly common, these issues are not trivial. Thus, it is important to work toward a clearer understanding when we do and do not have statistically significant crud in our results.

Random Crud

Few effect sizes are exactly $r = .00$, even for constructs most would agree have little relationship to each other. Thus, effect sizes can be expected to vary from .00 naturally to some degree, potentially on a normal curve. However, normal variation in effect sizes around 0 is poorly understood. That is to say, the mean effect sizes, standard deviation, and potential range of crud effect sizes have not been substantially investigated. If we accept that crud effect sizes vary and are dispersed around $r = .00$, we can expect that in large samples, many of these random but non-zero effect sizes will become statistically significant. Further, we would expect to see a bias toward positive crud. That is to say, given the existence of publication bias, authors, and journal editors are more likely to report and publish positive crud that supports an a priori hypothesis (van Assen et al., 2015). Aside from a bias toward positive crud, random crud is entirely beyond a researcher's control.

Non-Random Crud

Non-random crud occurs to the degree that methodological issues within psychological research are likely to nudge effect sizes toward the direction of a study's hypotheses. Such crud is non-random as it will tend to always preference the study hypotheses. Again, these nudges may be relatively small but, once again, in large sample studies may be enough to produce statistically significant results. Examples of phenomenon likely to cause non-random crud include demand characteristics or hypothesis guessing among participants (Sharpe & Whelton, 2016), single-responder bias in which predictor and outcome variables are both based on a single participants' survey responses (Baumrind et al., 2002), the use of unstandardized and unreliable measures (Elson et al., 2014) and researcher expectancy effects wherein researchers themselves unwittingly influence a participant's responses (MacCoun & Perlmutter, 2017). These biases can become systematic, replicated across studies in a field,

leading to artificial confidence in research results (Larzelere et al., 2015).

Non-random crud differs from random crud in several ways. First, it is at least partially under the control of the researcher, though perhaps impossible to eliminate entirely. Second and related, it can be reduced by certain procedures such as distractor items for demand characteristics, multiple-responder methods, and researcher blinding to condition. Third, the variance in non-random crud may be more difficult to estimate as it likely varies widely depending on the individual research methodology of particular studies.

The Current Study

Based on the previous literature, several questions are worth considering:

1. What is the potential for small effect sizes to be representative of methodological artifacts or random variance in effect sizes rather than real effects that exist in the population?
2. Is there any point to reporting effect sizes at all if they are all considered of evidentiary value so long as they cross $p < .05$?
3. Is it possible to estimate a high-risk threshold for effect sizes, wherein spurious correlations may be misinterpreted because they cross $p < .05$ in large samples?

The current study seeks to address these questions in a preliminary way using a large sample of adolescents. Given that random crud is likely more able to be estimated, the analyses will specifically seek to address some preliminary estimates for a lower-bound cutoff for effect sizes at high risk for spurious interpretation. Such an analysis may help scholars decide whether a tiny but statistically significant effect size can be safely interpreted as hypothesis supportive.

Disclosures

The preregistration for Study 1 is available at: https://osf.io/e9mtb/register/564d31db8c5e4a7c9694b2be. The code for Study 2 is presented at: https://osf.io/epfd5/.

Method

Participants

The current study makes use of two databases: (a) the National Longitudinal Study of Adolescent Health database (AddHealth; Resnick et al., 1997) and (b) data of the study Understanding Society, Wave 1 (2009–2011) Questionnaires, Youth Self-Completion (10–15 years) from the UK Data Archive (University of Essex, Institute for Social and Economic Research, 2019).

Study 1

This database is a large, publicly available (albeit for a fee), compilation of survey and in-person data with adolescents that is nationally representative. A full description of the methodology by which the database was developed can be found in Resnick et al. (1997). The current sample includes 20,403 adolescents who had completed information related to delinquency during the original assessment. Mean age was 15.65 ($SD = 1.74$) and 50.5% of the sample were female.

Study 2

The current sample includes 4,899 adolescents who had completed information related to conduct problems during the original assessment. This Understanding Society dataset is part of an ongoing study of UK households begun in 1991. The Understanding Society dataset specifically has been administered with adolescents every year from 2009 to 2017. Mean age was 12.51 ($SD = 1.70$) and 50.3% of the sample were female. A full description of the database is provided in University of Essex, Institute for Social and Economic Research (2019).

Measures

To study aggression the main outcome, a 15-item measure of self-reported delinquency was used. Sample items include "How often did you use or threaten to use a weapon to get something from someone?" and "How often did you drive a car without its owner's permission?" Omega hierarchical reliability for this subscale was .69.

For predictors, 20 crud items were initially chosen, 15 from either the survey self-responses or the in-person interview parts of the assessment, and five school variables. These are all indicated in the preregistration for this project. However, upon inspection, the five school variables were only recorded at the level of the school, not the individual and, as such, were not analyzed. All variables were chosen, similar to Orben and Przybylski (2019), for lack of theoretical rationale for why the predictors should predict delinquency. All variables are presented in the results section. These predictors typically involved either ordinal or yes/no responses. Variables were chosen from three waves of Study 1 as indicated by W1, W2, and W3, which stretched into adulthood.

Lastly, given that screen use was indicated as a potential predictor of negative mental health outcomes, including aggression, in much of the prior literature used as an example in this article, a measure of screen use was calculated. This measure combined two items measuring the frequency of self-reported television and computer use.

Study 2: We used the Subscale: Conduct Problems from the Strengths and Difficulties Questionnaire (SDQ). The SDQ is a five-item emotional and behavioral questionnaire for children and adolescents. Sample items include "I am often accused of lying or cheating," "I fight a lot. I can make other people do what I want," and "I get very angry and often lose my temper." See also www.sdqinfo.com for more information. Omega hierarchical reliability for this subscale was .64.

Procedure

Study 1

The procedure for this study was preregistered and this can be found at: https://osf.io/e9mtb/register/564d31db8c5e4a7c9694b2be. Both bivariate and partial r (controlling for gender) values were calculated for each of the 15 crud variables with delinquency.

Meta-analysis was also used to calculate a mean effect sizes from the absolute value of the correlation coefficients given there was no presupposed theoretical direction for crud in any case. Differences in mean effect size between ordinal and yes/no variables were also examined.

Study 2

This was not preregistered as it was not part of the initial study plan, but conducted in response to some initial comments on our manuscript. Thus, it may be fairly considered more exploratory than Study 1. However, it was intended as reproducibility study to examine whether the phenomenon-revealed in Study 1 were consistent across other datasets. Partial correlations controlling for age and gender were calculated between the 14 crud variables and conduct disorder symptoms as described in Table 2. The data are not publicly available but the analysis script written in R to rerun our analysis if one has access to the data can be found at https://osf.io/epfd5/. Furthermore, we generated five synthesized data sets of the original data using the R package synthpop (Nowok et al., 2016). These data do not contain the same values of the original data but retain the structure of the original data. Any inference therefore returns the same conclusion as the original. The synthesized data and the associated R analysis script for the meta-analysis carried in this study can also be downloaded from the OSF project page mentioned above.

Results

Study 1

Both the bivariate and partial *r* correlations between the 15 crud variables as well as screen use and self-reported delinquency are reported in Table 1. Results showed general consistency across methods variance, including Pearson *r,* Spearman rho, and partial *r*. In each case, of the 15 crud variables, between 7 and 10 of the variables were statistically significant. Highest effect sizes were seen for two crud variables in particular, frequency of sore throats and frequency of sunscreen use. These effect sizes were actually larger than for screen use, a variable of interest to many scholars.

Meta-analysis was run using the Comprehensive Meta-Analysis program. Mean fixed effects models were consistent at between $r = .033$ and .035 across correlation methods. Effect sizes were higher for ordinal items ($r = .043$ in bivariate analyses) than yes/no items ($r = .021$).

Study 2

Because all crud variables were unordered categorical variables, we ran simple linear regressions by regressing the scores of the subscale Conduct Problems on each of the variables (see Table 2) to estimate *R*, that is, the correlation between the predicted and observed conduct problem scores. To furthermore obtain partial correlations between the unordered categorical crud variables and conduct problem scores, controlled for "age" and "gender," we took the square root of estimated partial omega coefficients. This approach is derived from Levine and Hullett (2002, p. 622) who argue that partial eta squared, which is the upward-biased counterpart of partial omega squared (Maxwell & Delaney, 1990) is conceptually equivalent to the squared partial correlation. It should be noted that in the case of the crud variable "interview date," partial omega squared turned out to be slightly negative (−.001) and the square root could therefore not be calculated. Negative estimates of partial omega squared can occur because the sampling distribution of omega squared can have a substantial mass in the negative region when population effects are small (Okada, 2017). Such negative estimates do, however, not make sense because partial omega squared as the ratio of explained variability to variability unexplained by all the other predictors of a model must lie between 0 and 1. This negative estimate could therefore neither be used in the significance tests nor in the meta-analysis mentioned below, thereby reducing the number of results being part of meta-analysis from 14 to 13.

Table 1
Correlations Between Crud and Delinquency for Study 1

Crud variable	*r*	rho	Partial *r*
Sore Throat Frequency W1	.093***	.106***	.102**
Asthma W1	.013	.005	.019
Driving, Miles/Week W1	−.016*	−.023**	−.026**
Plays a Sport W2	.017*	.016	−.021*
Drinks Milk W2	−.018*	−.010	−.034***
Drinks Water W2	−.018*	−.021*	−.025**
Frequency Sunscreen Use W2	.098***	.107**	.074***
Past Year Medical Exam W2	.005	.007	−.002
Respondent Lives at Interview Site W3	−.001	−.012	−.001
Bedtime Hour W3	−.043***	−.032***	−.034***
Does Work Around House W3	−.007	−.004	.029***
Adoption Status W1	.043***	.034***	.040***
Time Since Previous Dental Exam W1	.016	.011	.013
Uses Artificial Limb W1	.041***	.029**	.029**
Age of Child W3	.009	.008	.019
Screen Use W1	.084***	.084***	.069**

Note. W = study wave. Partial *r* control for gender.
* $p = .05$. ** $p < .01$. *** $p < .001$.

The significance of the squared partial omega coefficients was determined using bootstrapped 95% confidence intervals with standard errors being based on the sample quantiles of the bootstrapped values. Eight out of 13 coefficients turned out to be significant. The results are displayed in Table 2.

Square-rooted partial omega square coefficients (yielding partial correlations) were then meta-analyzed to obtain an overall effect size measure using the R package metafor (Viechtbauer, 2010). A fixed effects meta-analysis yielded a significant mean effect size with $r_{partial} = .11$, $z = 26.11$, $p < .001$, two-tailed $CI_{95\%} = .098, .113$. (The five meta-analyses based on the five synthesized data yield an overall effect size of $r_{partial} = .11$, $SD = .01$). The results from the synthesized data aimed to provide reproducible data analyses are therefore highly comparable to those obtained from the original data set.)

Discussion

Several decades ago Meehl (1991) suggested that many small findings in psychology could be "crud," a tendency for everything to correlate with everything else a tiny amount. The current article sought to provide some preliminary examination of the crud phenomenon-in a large sample of adolescents with aggressive

Table 2
Correlations Between Crud and Conduct Problems for Study 2

Crud variable	Number of valid response options used	R	$\sqrt{\omega^2_{partial}}$
Siblings at home	2	.037**	.034 *ns*
Head-aches, stomach-aches, or sickness	3	.222***	.224*
One good friend or more	3	.050***	.049 *ns*
What would most like to do at 16	5	.180***	.178*
How often eat crisps fizzy drinks sweets	4	.138***	.139*
Main means of travel to school	6	.043 *ns*	.029 *ns*
Religious membership in Great Britain	10	.088***	.076*
Sex of natural parent with lowest personal number	2	.052***	.051 *ns*
Job would like when left education based on Standard Occupational Classification 1990	371	.266***	.181*
Job would like when left education based on Standard Occupational Classification 2000	353	.278***	.186*
Job would like when left education based on Standard Occupational Classification 2010	369	.277***	.179*
Ethnic group	18	.073 *ns*	.044*
Interview date: day	31	.075 *ns*	–[a]
Country of residence	4	.026 *ns*	.007 *ns*

Note. The significance of $\sqrt{\omega^2_{partial}}$ is based on the significance of partial omega square values.
[a] Could not be computed because omega square was negative.
* $p = .05$. ** $p < .01$. *** $p < .001$.

delinquency as outcome. Analyses from Study 1 suggest that with effect sizes below $r = .10$, a majority of nonsense relationships achieve statistical significance, with some approaching or slightly exceeding the value of $r = .10$. These results suggest a higher than tolerable probability for false positive findings among effect sizes below $r = .10$ among large sample studies. It is interesting that some of these variables produced effect sizes higher than for screen use, a variable often considered important by psychologists when considering aggression. This highlights the potential for spurious interpretation of weak effects.

With Study 2, mean effect sizes were slightly higher, around $r = .11$. This suggests that even some effect sizes exceeding $r = .11$ and "statistically significant" in large datasets may be artifactual in nature. Though there is no definite hard "cut-off" effect size, confidence in the meaningfulness of statistically significant effect sizes should clearly decrease the nearer they approximate 0. In the two datasets, only a single crud relationship exceeded $r = .20$.

As a practical suggestion, it is recommended that effect sizes below $r = .10$ should not be interpreted as evidence in support of a hypothesis, at least of one claiming the existence of a univariable relationship or a univariable causal effect. Such weak data are likely inconclusive at best or may just reflect a hodgepodge of relationships or causal effects of more than one variable. In the past, many scholars have constructed arguments for why tiny effects may nonetheless be important. For instance, scholars have sometimes suggested that tiny effects spread across a population can nonetheless have practically significant impact, or that important medical findings sometimes have tiny effect sizes in terms of *r*. These latter arguments appear to have been discredited as mainly due to statistical calculation errors (Ferguson, 2009) whereas the former extrapolates within-participant variance to populations in an inappropriate way. Moreover, although correlations of such a magnitude may represent something "real" in terms of the aforementioned hodgepodge- or net-effects/relationships among variables, such effects might just be either too small to care about in light of more obvious contributors or inconclusive because they cannot be interpreted as relationships between just two variables. Thus, it is recommended that psychological science adopt more conservative standards for the interpretation of tiny or small effect sizes. If effect sizes are a game of all have won and must have prizes, reporting them in the first place is moot.

It is possible that, in some cases, scholars may have valid reasons for concluding that an effect sizes might be truncated due to issues such as unreliable measures. This should not be used as an argument for interpreting the observed effect sizes as hypothesis supportive. However, authors could provide recommendations for how future studies could examine the issue further using more precise techniques if there are concerns in this regard.

It is noted that effect sizes above $r = .10$ are no guarantee for having escaped "crud." Non-random crud could be much larger due to methodological limitations, including systematic methodological limitations across particular fields. The current study applies only to random crud, not non-random crud. It is likely that many false positive results exceed $r = .10$. However, $r = .10$ appears to be a reasonable minimal cut-off for a likely signal versus noise problem in psychological research, with the understanding that noise and crud effects may actually extend far higher (e.g., Meehl, 1990, 1997; Waller, 2004). Effects between $r = .10$ and $r = .20$ may also be regarded with caution as the proportion of uninterpretable results here are likely to remain fairly high (Lykken, 1968). As such, the .10 cutoff should be regarded as basement cut-off under which an effect should not be interpreted as hypothesis supportive. However, exceeding .10 or even .20 is not a guarantee that an effect is "real" and various noise or crud factor issues may still lead to misinterpretations of the importance of particular findings.

Naturally, use of any hard and fast cut-off brings with it some limitations. It is important that scholars focus on theoretically relevant predictors and reproducibility. However, it is entirely

possible that a theoretically hypothesized but tiny effect may appear reproducible in large datasets, but this being due to systematic methodological problems in the field rather than a "true" effect. As such, we do believe that a threshold for interpretation as hypothesis supportive, below which we know many effect sizes are spurious, will be helpful for researchers interested in avoiding overinterpretation of weak and potentially spurious data.

Limitations

Although the current analyses employed a rigorously developed dataset with two large, nationally representative samples, it is necessarily limited in scope. In particular, the current analyses test crud in aggression research. Crud in other fields of research may be larger or smaller, although it is unlikely to ever be absent. Ultimately, the current figures are only a preliminary estimate and more work on this issue would be welcome. The current results should not be used to warn parents of the dangers of sore throats or sunscreen as risk factors for aggression.

In our study, "nonsense" variables were chosen for their theoretical lack of relationship to aggression. Another approach would have been to simply include a random selection of variables against which to correlate aggression. Both approaches have their value, however we thought it is best to use "nonsense" variables to get a clearer picture of noise effects from which true signal could be distinguished. Had we employed a random variable sample, some of those variables would have been theoretically relevant for aggression. This would have increased the effect size as this calculated effect would have included both "true" effects and "crud" effects. In our view this resultant effect size would not have been a conservative estimate of "crud" and could have resulted in potential rejection of some "true" (albeit small) effects. As we wished to provide an estimate of crud that was noise or nonsense only, we did not take this approach. We certainly understand that there is risk of variable selection bias in our approach, but a random sampling approach would work best only in a dataset from which the majority of variables could be expected to be unrelated.

Concluding Thoughts

In recent years, psychological science has struggled with a replication crisis that has challenged many previously held truisms (Pashler & Harris, 2012). An overreliance on tiny effect sizes may also be promoting many false positive results, even when issues such as p-hacking or other questionable researcher practices are not in play. For psychological science to become surer in its findings, adoption of a higher threshold of evidence will likely be necessary. This will undoubtedly require abandoning many trivial effect sizes that are "statistically significant" but, nonetheless, crud.

References

Adachi, P. J. C., & Willoughby, T. (2011). The effect of video game competition and violence on aggressive behavior: Which characteristic has the greatest influence? *Psychology of Violence*, *1*(4), 259–274. https://doi.org/10.1037/a0024908

Baumrind, D., Larzelere, R. E., & Cowan, P. A. (2002). Ordinary physical punishment: Is it harmful? Comment on Gershoff (2002). *Psychological Bulletin*, *128*(4), 580–589. https://doi.org/10.1037/0033-2909.128.4.580

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*(2), 419–432. https://doi.org/10.1037/a0035569

Ferguson, C. J. (2009). Is psychological research really as good as medical research? Effect size comparisons between psychology and medicine. *Review of General Psychology*, *13*(2), 130–136.

Furuya-Kanamori, F., & Doi, S. (2016). Angry birds, angry children and angry meta-analysts. *Perspectives on Psychological Science*, *11*(3), 408–414.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*(11), 1337–1345.

Heene, M. (2020, December 9). *AddH*. osf.io/epfd5

Heffer, T., Good, M., Daly, O., McDonnell, E., & Willoughby, T. (2019). The longitudinal association between social-media use and depressive symptoms among adolescents and young adults: An empirical reply to Twenge et al. (2018). *Clinical Psychological Science*, *7*(3), 462–470.

Huesmann, L. R. (2007). The impact of electronic media violence: Scientific theory and research. *Journal of Adolescent Health*, *41*, S6–S13.

Larzelere, R. E., Cox, R. B., Jr., & Swindle, T. M. (2015). Many replications do not causal inferences make: The need for critical replications to test competing explanations of nonrandomized studies. *Perspectives on Psychological Science*, *10*(3), 380–389. https://doi.org/10.1177/1745691614567904

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research*, *28*(4), 612–625. https://doi.org/10.1111/j.1468-2958.2002.tb00828.x

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*(3, Pt.1), 151–159. https://doi.org/10.1037/h0026141

MacCoun, R. J., & Perlmutter, S. (2017). Blind analysis as a correction for confirmatory bias in physics and in psychology. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 297–322). Wiley-Blackwell.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data*. Brooks Cole.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Meehl, P. E. (1991). Why summaries of research on psychological theories are often uninterpretable. In R. E. Snow, D. E. Wiley, R. E. Snow, & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 13–59). Erlbaum.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393–425). Erlbaum.

Norcross, J. C., Hogan, T., Koocher, G. P., & Maggio, L. A. (2017). *Clinician's guide to evidence-based practices: Mental health and the addictions* (2nd ed.). Oxford University Press.

Nowok, B., Raab, G. M., & Dibben, C. (2016). Synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, *74*(11), 1–26. https://doi.org/10.18637/jss.v074.i11

Okada, K. (2017). Negative estimate of variance-accounted-for effect size: How often it is obtained, and what happens if it is treated as zero. *Behavior Research Methods*, *49*(3), 979–987. https://doi.org/10.3758/s13428-016-0760-y

Orben, A., Deinlin, T., & Przybylski, A. (2019). Social media's enduring effect on adolescent life satisfaction. *Proceedings of the National Academy of Sciences*, *116*(21), 10226–10228.

Orben, A., & Lakens, D. (2020). Crud (re)defined. *Advances in Methods and Practices in Psychological Science*, *3*(2), 238–247. https://doi.org/10.1177/2515245920917961

Orben, A., & Przybylski, A. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, *3*, 173–182.

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

Prescott, A. T., Sargent, J. D., & Hull, J. G. (2018). Metaanalysis of the relationship between violent video game play and physical aggression over time. *PNAS: Proceedings of the National Academy of Sciences of the United States of America*, *115*(40), 9882–9888. https://doi.org/10.1073/pnas.1611617114

Redding, R. (2001). Sociopolitical diversity in psychology: The case for pluralism. *American Psychologist*, *56*, 205–215.

Resnick, M. D., Bearman, P. S., Blum, R. M., Bauman, K. E., Harris, K. M., Jones, J., Tabor, J., Beuhring, T., Sieving, R. E., Shew, M., Ireland, M., Bearinger, L. H., & Udry, J. (1997). Protecting adolescents from harm: Findings from the National Longitudinal Study on Adolescent Health. *Journal of the American Medical Association*, *278*(10), 823–832. https://doi.org/10.1001/jama.278.10.823

Sharpe, D., & Whelton, W. J. (2016). Frightened by an old scarecrow: The remarkable resilience of demand characteristics. *Review of General Psychology*, *20*(4), 349–368. https://doi.org/10.1037/gpr0000087

Standing, L., Sproule, R., & Khouzam, N. (1991). Empirical statistics: IV. Illustrating Meehl's sixth law of soft psychology: Everything correlates with everything. *Psychological Reports*, *69*(1), 123–126. https://doi.org/10.2466/PR0.69.5.123-126

Twenge, J. M., Joiner, T. E., Rogers, M. L., & Martin, G. N. (2018). Increases in depressive symptoms, suicide-related outcomes, and suicide rates among US adolescents after 2010 and links to increased new media screen time. *Clinical Psychological Science*, *6*(1), 3–17. https://doi.org/10.1177/2167702617723376

University of Essex, Institute for Social and Economic Research. (2019). *Understanding Society: Waves 1-9, 2009–2018 and Harmonised BHPS: Waves 1-18, 1991–2009. [data collection] (12th ed.)*. UK Data Service. SN: 6614, https://doi.org/10.5255/UKDA-SN-6614-13

Vachon, D. D., Lynam, D. R., & Johnson, J. A. (2014). The (non)relation between empathy and aggression: Surprising results from a meta-analysis. *Psychological Bulletin*, *140*(3), 751–773. https://doi.org/10.1037/a0035236

van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, *20*(3), 293–309. https://doi.org/10.1037/met0000025

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://www.jstatsoft.org/v36/i03/

Waller, N. (2004). The fallacy of the null hypothesis in soft psychology. *Applied & Preventive Psychology*, *11*, 83–86.

Wilkinson, L., & Task Force on Statistical Inference, American Psychological Association, Science Directorate. (1999). Statistical methods in psychological journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Received August 28, 2020
Revision received December 9, 2020
Accepted February 12, 2021 ■