This manuscript has been accepted for publication at Perspectives on Psychological Science

Are Small Effects the Indispensable Foundation for a Cumulative Psychological Science? A Reply to Götz et al. (2022)

Maximilian A. Primbs*[1], Charlotte R. Pennington[2,3], Daniël Lakens[4], Miguel Alejandro A. Silan[5,6,7], Dwayne S. N. Lieck[8], Patrick S. Forscher[9], Erin M. Buchanan[10], Samuel J Westwood[11]

*Corresponding author.

[1]Behavioural Science Institute, Radboud University, Nijmegen, the Netherlands.
max.primbs@ru.nl, https://orcid.org/0000-0002-3398-5569

[2] School of Psychology, Aston University, Birmingham, UK.
c.pennington@aston.ac.uk, https://orcid.org/0000-0002-5259-642X

[3] Institute of Health & Neurodevelopment, Aston University.

[4] School of Innovation Sciences, Eindhoven University of Technology, The Netherlands.
D.Lakens@tue.nl, https://orcid.org/0000-0002-0247-239X

[5] Annecy Behavioral Science Lab, France.
MiguelSilan@gmail.com , https://orcid.org/0000-0002-7480-3661

[6] Development, Individual, Process, Handicap, Education (DIPHE) Research Unit, Université Lumière Lyon 2.

[7] Social and Political Psychology Research Lab, University of the Philippines Diliman.

[8] Independent Researcher.
dwayne.lieck@gmail.com, https://orcid.org/0000-0001-9814-3488

[9] Busara Center for Behavioral Economics, Nairobi, Kenya.
schnarrd@gmail.com, https://orcid.org/0000-0002-7763-3565

[10] Harrisburg University of Science and Technology, Analytics, USA.
ebuchanan@harrisburgu.edu, https://orcid.org/0000-0002-9689-4189

[11] Department of Psychology, School of Social Science, University of Westminster, London, UK. s.westwood@westminster.ac.uk
https://orcid.org/0000-0002-0107-6651

Authorship Contributions: Conceptualization - MP, CRP, DL, MAAS, DSNL, PF, EMB; SJW. Writing - Original Draft - MP, CRP, DL., MAAS, DSNL, PF, EMB; Writing - Review & Editing: MP, CRP, SJW, DL, MAAS, DSNL, EMB, PF; Supervision - MP, CRP, DL; Project administration - MP, CRP.

Declaration of Conflicting Interests

The authors do not declare any conflicts of interest.

Funding

Daniël Lakens was funded by the Netherlands Organisation for Scientific Research VIDI Grant 452-17-013.

Acknowledgements: The authors would like to thank Stuart J. Ritchie, Jonathan Coleman and Freek Oude Maatman for their comments on earlier drafts of the section titled '*Genetics is not a Useful Analogy for Psychology'.*

Abstract

Götz et al. (2022) argue that small effects are “the indispensable foundation for a cumulative psychological science”. They support their argument by claiming that (i) psychology, like genetics, consists of complex phenomena explained by additive small effects, (ii) psychological research culture rewards large effects, which means small effects are being ignored, and (iii) small effects become meaningful at scale and over time. We rebut these claims with three objections: (i) the analogy between genetics and psychology is misleading, (ii) *p*-values are the main currency for publication in psychology, meaning that any biases in the literature are (currently) caused by a pressure to publish statistically significant results and not large effects, and (iii) claims regarding small effects as important and consequential must be supported by empirical evidence or, at least, a falsifiable line of reasoning. If accepted uncritically, we believe the arguments of Götz et al. (2022) could be used as a blanket justification for the importance of *any* and *all* ‘small’ effects, thereby undermining best practices in effect size interpretation. We end with guidance on evaluating effect sizes in relative, not absolute terms.

Key words: effect sizes; small effects; benchmarks; practical significance; statistical inference

Are Small Effects the Indispensable Foundation for a Cumulative Psychological Science? A Reply to Götz et al. (2022)

In their recent commentary, Götz et al. (2022) argue that small effects are "the indispensable foundation for a cumulative psychological science". Although we welcome their efforts to highlight the importance of reporting and interpreting effect sizes appropriately, we believe that some of their arguments have the potential to move us away from, and not towards, best practices in effect size interpretation. Here we counter their arguments with three objections: (i) the analogy between genetics and psychology is misleading, (ii) selection for statistical significance (*p*-values) rather than selection for large effect sizes currently underpins publication bias, and (iii) statements that small effects may be important and consequential need to be supported by evidence and falsifiable reasoning rather than bald assertion. Furthermore, we disagree with Götz et al.'s assumption that 'small' and 'large' are meaningful categories outside of a particular theoretical and empirical context. We argue that effect sizes should be interpreted in relative and not absolute terms.

Genetics is not a useful analogy for psychology.

Götz et al. (2022) begin their line of reasoning by drawing an analogy between psychology and genetics: they argue that like the links between genes and behaviour, psychological phenomena have multiple complex causal mechanisms, which means that the effects of any single mechanism are bound to be small. Although we broadly agree with the notion that psychological phenomena have multiple complex causal mechanisms, we are agnostic about the size of individual effects. Further, we argue that the analogy between genetics[1] and psychology can confuse, more than progress, knowledge accumulation in psychological science for two main reasons.

[1] We use the term genetics because it has been used by Götz et al.; We recognize that the practices outlined by Götz et al. more resemble those used in polygenic modelling of genome-wide association studies and may not be applicable to other areas of genetics.

First, in the study of genetics, there are a known and finite set of genes that can be measured accurately and quickly (see the Human Genome Project, Schmutz et al., 2004), but the same cannot be said for psychological phenomena. Even if it were possible to list all constructs of interest, it is not feasible to measure all of the constructs that psychologists might be interested in testing. In addition, psychologists are often interested in theoretically driven tests of predictions, which typically require specific experimental manipulations, and often involve tests of interaction effects. Open datasets that are large enough to provide sufficient statistical power to detect small effects at a similar scale to genetics do not currently exist, and it is unlikely that psychologists will have the resources to create such datasets. This makes "Psychological Construct Association Studies" (to continue the analogy by Götz et al.) an impossible approach that will not enable researchers to efficiently generate knowledge about complex psychological phenomena.

Second, the scales used to measure psychological phenomena are much more coarse-grained than the measurement of genes, with many lacking the accuracy to reliably detect small effects (Flake et al., 2017; Fried, 2017). Until our measurement practices have improved, psychologists will not be able to distinguish measurement error from small effects. In reality, this makes it impossible to reliably study the small effects that Götz and colleagues hypothesise to be the indispensable foundation of a cumulative psychological science. Furthermore, even if measurement accuracy were perfect, we would still need to address the challenge of reliably distinguishing small effects of interest from 'crud' — the notion that in large enough datasets in psychological science, all variables are correlated with each other (Meehl, 1990; Orben & Lakens, 2020; Vul et al., 2009). This is especially challenging because in some domains the crud factor is hypothesised to be as large as $r = 0.10$ (Ferguson, 2021).

In sum, we argue that caution is warranted when using genetics as an analogy for psychological science: given the current state of affairs in psychological science, such as measurement imprecision, we simply do not know whether psychological phenomena are indeed caused by many additive small effects. One could even argue that given the limits of

human information processing, small effects may not matter because they are simply not perceived.

P-values (not effect sizes) currently underpin publication bias and questionable research practices

Contrary to what Götz et al. (2022) report, Fanelli et al. (2017) do not find that "Social scientific disciplines often cultivate publication cultures that favor or even demand large effects" (p.206). Instead they show that small studies can overestimate effect sizes and that early studies in some fields have larger effects. Neither of these findings show any demand for large effects, but instead show the limitations of underpowered studies that lead to inflated, unreliable effect sizes. Instead, publication bias is underpinned by a preoccupation with *p*-values: effects that are statistically significant are published at a higher rate than non-significant effects in the traditional literature (see Fanelli, 2010; Scheel et al., 2021). In support of this, research shows that researchers often do not interpret effect sizes (Fritz et al., 2013; Motyl et al., 2017; Schäfer & Schwarz, 2019), and when requested by reviewers or editors to do so, this is usually on the basis of *justifying* whether certain "significant" effects matter, rather than dismissing small effects altogether.

Moreover, Götz et al. (2022) argue that "the pressure to publish large effects is "dangerous" because it (...) encourages practices that are likely to yield these inflated effects such as *p*-hacking, optional stopping, HARKing, and other questionable research practices [QRPs]" (p. 206). The smallest effect size that corresponds to a statistically significant result is a function of the alpha level and the sample size. Given a tradition of running small underpowered studies and selectively reporting statistically significant results (see Button et al., 2013; Szucs & Ioannidis, 2017), the mechanism Götz et al. describe is, in fact, reversed: it is not a pressure to publish large effects that encourages QRPs, but rather QRPs coupled with low statistical power that inflate effect sizes to reach "publishable" p-values (see Stefan & Schönbrodt, 2022). Whilst we wholeheartedly agree with Götz et al. (2022) that effect sizes are important, and should be evaluated in terms of their theoretical and practical

applications, we feel that it is important to correct the basis of some of their claims: there is currently no empirical support to suggest that large effects are favoured or demanded.

Claims that small effects can be important and consequential requires empirical evidence.

Götz et al. (2022) state further that “some small effects may also have direct real-world consequences (Funder & Ozer, 2019; Gelman & Carlin, 2014). This phenomenon is especially true for effects that accumulate over time and at scale” (p. 206). To support this claim, Götz et al. cite research on the Implicit Association Test (IAT; Greenwald et al., 2015), which claims precisely such accumulation. However, IAT researchers have been unable to provide empirical evidence for this accumulation and do not theoretically specify how such accumulation may occur (Connor & Evers, 2020). Any argument for the accumulation of a ‘small’ effect must therefore be substantiated by empirical evidence rather than speculation or, at least, be supported by a falsifiable line of reasoning. This should also consider any possible mechanisms that may act against such accumulation (e.g., habituation; Anvari et al., 2021; Funder & Ozer, 2019) as well as those that facilitate it.

To further illustrate the claim that small effects can be consequential in large samples or at the population level, Götz et al. (2022) present the correlation between aspirin and the prevention of heart attacks (r = .03). Ferguson (2009) has pointed out the flaw in using this effect size to make the generalised argument that small effects matter through an analogy with wearing a bulletproof vest: the effect size of wearing a bulletproof vest on the probability of dying is large if we examine people who get shot, but very small if we include millions of people who never get shot. Likewise, the causal effect of aspirin on the chance of a heart attack is substantial, but there is only a small effect in the reduction of heart attacks if a large group of people, many of which would never suffer a heart attack, regularly take aspirin. The important difference between medicine and psychology is that in psychology, researchers rarely include a large majority of individuals in their studies that are not expected to benefit from an intervention. For example, when we examine the effectiveness of a new treatment

for depression, we usually do not conduct the study on a sample where only a small minority of individuals are depressed. Therefore, when a small effect is observed in psychology, it may not matter at the population level and any claims of why it would matter need to be theoretically justified and/or empirically supported.

Categorizing r = .03 as 'small' regardless of empirical context, discipline, study design, and outcome variable is, in short, nonsensical. If an intervention saves hundreds of thousands of lives, then its effect on human health and society is by no reasonable definition 'small'. If the cost of the intervention is as low as an aspirin, it is likely worthwhile to implement in practice[2]. To judge an effect as meaningful, one needs to provide evidence and a line of falsifiable reasoning.

Summary and discussion

Götz et al. (2022, pp. 205) state that "only once small effects are accepted as the norm, rather than the exception, can a reliable and reproducible cumulative psychological science be built". They claim that (i) psychology, like genetics, consists of complex phenomena explained by additive small effects, (ii) psychology should not only reward large effects, and (iii) small effects become meaningful at scale and over time. In this reply, we present counter-arguments outlining (i) that we cannot currently make claims about the size of effects influencing psychological phenomena in the same way as genetics, (ii) that statistical significance and not effect sizes underpin publication bias and QRPs, and (iii) that claims that small effects are important at scale or over time must be supported by empirical evidence and a falsifiable line of reasoning. We suggest that researchers must evaluate the meaningfulness of an effect size in respect to its theoretical and empirical context.

We argue that researchers should move away from interpreting effect sizes in an absolute manner: that is, there are no 'small' or 'large' effects in isolation of their contextual

[2] But see https://www.uspreventiveservicestaskforce.org/uspstf/announcements/public-comment-draft-recommendation-statement-draft-evidence-review-and-draft-modeling-report-aspirin-use-prevent-cardiovascular. Further, as pointed out by Robert Calin-Jageman, the effect is more interpretable expressed as an odds ratio.

factors. Researchers should therefore adopt a *relative* framework to effect size interpretation, where the size of an effect is compared to its costs (i.e., practical or substantive significance, Kelley & Preacher, 2012; Silan, 2020), other effects in the same empirical context (e.g., this treatment effect is larger than effect sizes of other treatments), or a benchmark such as the smallest effect size of interest or maximal positive control that is established through appropriate empirics, theory, or falsifiable justification (see Anvari & Lakens, 2021; Hilgard, 2021; Rocca & Yarkoni, 2021).

Ultimately, statements about effect sizes cannot be reduced to a mechanical process and researchers need to provide arguments that support why any effect, of any size, should be considered relevant. As psychologists start to collect larger sample sizes, and restrict flexibility in their statistical analyses through the adoption of open science practices, they will observe more accurate effect size estimates. We are concerned that researchers confronted with very small but statistically significant effect size estimates will cite Götz et al. (2022) as a blanket defence for why *any* or *all* small effects matter, and indeed, we are already witnessing signs of this see Dickey et al., 2021; Greenberg et al., 2022; Jokela, 2021; Rimfeld et al., 2021; Sorlie et al., 2022). Instead, we urge researchers to justify their effect sizes and to think about the practical significance of these effects; a practice that is likely to differ between disciplines and research fields.

References

Anvari, F., Kievit, R., Lakens, D., Pennington, C. R., Przybylski, A. K., Tiokhin, L., Wiernik, B. M., & Orben, A. (2021). Evaluating the practical relevance of observed effect sizes in psychological research. https://psyarxiv.com/g3vtr/

Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology, 96,* 104159. https://doi.org/10.1016/j.jesp.2021.104159

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 365-376. https://doi.org/10.1038/nrn3475

Connor, P., & Evers, E. R. K. (2020). The bias of individuals (in crowds): Why implicit bias is probably a noisily measured individual-level construct. *Perspectives on Psychological Science, 15*(6), 1329-1345. https://doi.org/10.1177/1745691620931492

Dickey, L., West, M., Pegg, S., Green, H., & Kujawa, A. (2021). Neurophysiological responses to emotional images prospectively predict the impact of Covid-19 pandemic-related stress on internalising symptoms. *Biological Psychiatry, 6*(9), 887-897. https://doi.org/10.1016/j.bpsc.2021.03.004

Fanelli, D. (2010). “Positive” results increase down the hierarchy of the sciences. *PLoS One, 5*(4), e10068. https://doi.org10.1371/journal.pone.0010068

Fanelli, D., Costas, R., & Ioannidis, J. P. A. (2017). Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences, 114*, 3714–3719. https://doi.org/10.1073/pnas.1618569114

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice, 40*(5), 532–538. https://doi.org/10.1037/a0015808

Ferguson, C. J. (2021). Providing a lower-bound estimate for psychology's "crud factor": The case of aggression. *Professional Psychology: Research and Practice.* http://dx.doi.org.dianus.libr.tue.nl/10.1037/pro0000386

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

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191-197. https://doi.org/10.1016/j.jad.2016.10.019

Fritz, A., Scherndl, T., & Kühberger, A. (2013). A comprehensive review of reporting practices in psychological journals: Are effect sizes really enough? *Theory & Psychology*, *23*(1), 98-122. https://doi.org/10.1177/0959354312436870

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advanced in Methods and Practices in Psychological Science, 2*(2), 156-168. https://doi.org/10.1177/2515245919847202

Götz, F. M., Gosling, S. D., & Rentfrow, P. J. (2022). Small effects: The indispensable foundation for a cumulative psychological science. *Perspectives on Psychological Science*, *1,* 205-215.. https://doi.org/10.1177/1745691620984483

Greenberg, D. M., Wride, S. J., Snowden, D. A., Spathis, D., Potter, J., & Rentfrow, P. J. (2022). Universals and variations in musical preferences: A study of preferential reactions to Western music in 53 countries. *Journal of Personality and Social Psychology, 122*(2), 286–309. https://doi.org/10.1037/pspp0000397

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality & Social Psychology, 108*, 553–561. https://doi.org/10.1037/pspa0000016

Hilgard, J. (2021). Maximal positive controls: A method for estimating the largest plausible effect size. *Journal of Experimental Social Psychology*, *93*, 104082. https://doi.org/10.1016/j.jesp.2020.104082

Jokela, M. (2021). Urban-rural residential mobility associated with political party affiliation: The U.S. National Longitudinal Surveys of Youth and Young Adults. *Social Psychological & Personality Science, 13*(1), 83-90. https://doi.org/10.1177/1948550621994000

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137–152. https://doi.org/10.1037/a0028086

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. Psychological Reports, 66, 195-244. Retrieved from: https://meehl.umn.edu/sites/meehl.umn.edu/files/files/144whysummaries.pdf

Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., Prims, J. P., Sun, J., Washburn, A. N., Wong, K. M., Yantis, C., & Skitka, L. J. (2017). The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of Personality and Social Psychology, 113*(1), 34–58. https://doi.org/10.1037/pspa0000084

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

Rimfeld, K., Malanchini, M., Arathimoas, R., Gidziela, A., Pain, O., …, & Plomin, R. (2021). The consequences of a year of the Covid-19 pandemic for the mental health of young adult twins in England and Wales. Retrieved from: https://www.medrxiv.org/content/medrxiv/early/2021/10/07/2021.10.07.21264655.full.pdf

Rocca, R., & Yarkoni, T. (2021). Putting psychology to the test: Rethinking model evaluation through benchmarking and prediction. *Advances in Methods and Practices in Psychological Science, 4*(3), 1-24. https://doi.org/10.1177/25152459211026864

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology*, *10*, 813. https://doi.org/10.3389/fpsyg.2019.00813

Scheel, A. M., Schijen, M. R., & Lakens, D. (2021). An excess of positive results: Comparing the standard Psychology literature with Registered Reports. *Advances in Methods*

and Practices in Psychological Science, 4, 1-12. https://doi.org/10.1177/2515245921007467

Schmutz, J., Wheeler, J., Grimwood, J., Dickson, M., Yang, J., Caoile, C., Bajorek, E., Black, S., Chan, Y. M., Denys, M., Escobar, J., Flowers, D., Fotopulos, D., Garcia, C., Gomez, M., Gonzales, E., Haydu, L., Lopez, F., Ramirez, L., … Myers, R. M. (2004). Quality assessment of the human genome sequence. *Nature, 429*(6990), 365-368. https://doi.org/10.1038/nature02390

Silan, M. A. (2020, December 23). A Primer on Practical Significance. https://doi.org/10.31234/osf.io/zdhfe

Sorlie, H. O., Hetland, J., Bakker, A. B., Espevik, R., & Olsen, O. K. (2022). Daily autonomy and job performance. Does person-organization fit act as a key resource? *Journal of Vocational Behaviour, 133*, 103691. https://doi.org/10.1016/j.jvb.2022.103691

Stefan, A., & Schönbrodt, F. D. (2022, March 16). Big little lies: A compendium and simulation of p-hacking strategies. https://doi.org/10.31234/osf.io/xy2dk

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. https://doi.org/10.1371/journal.pbio.2000797

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies on emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*(3), 274-290.