Not all effects are indispensable: Psychological science requires verifiable lines of reasoning for whether an effect matters

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

Farid Anvari[1,2], Rogier A. Kievit[3], Daniël Lakens[4], Charlotte R. Pennington[5], Andrew K. Przybylski[6], Leo Tiokhin[4], Brenton M. Wiernik[7], and Amy Orben[8]

Abstract

To help move researchers away from heuristically dismissing "small" effects as unimportant, recent articles have revisited arguments to defend why seemingly small effect sizes in psychological science matter. One argument is based on the idea that an observed effect size may increase in impact when generalized to a new context due to processes of accumulation over time or application to large populations. However, the field is now in danger of heuristically accepting all effects as potentially important. We aim to encourage researchers to think thoroughly about the various mechanisms that may both amplify *and* counteract the importance of an observed effect size. Researchers should draw on the multiple amplifying and counteracting mechanisms that are likely to simultaneously apply to the effect when that effect is being generalized to a new (and likely more dynamic) context. In this way, researchers should aim to transparently provide verifiable lines of reasoning to justify their claims about an effect's (un)importance. This transparency can help move psychological science towards a more rigorous assessment of when psychological findings matter for the contexts that researchers want to generalize to.

Keywords

effect size, practical significance, benchmarks, evaluation

[1]Social Cognition Center Cologne, University of Cologne, Germany
[2]Strategic Organization Design group, University of Southern Denmark, Denmark
[3]Cognitive Neuroscience Department, Radboud University Medical Center, Nijmegen, The Netherlands
[4]Eindhoven University of Technology, The Netherlands
[5]School of Health and Life Sciences, Aston University, United Kingdom
[6]University of Oxford, The United Kingdom
[7]Department of Psychology, University of South Florida, Tampa, FL, USA
[8]University of Cambridge, The United Kingdom
Corresponding Author:
Farid Anvari. Email: farid.anvari@uni-koeln.de.

To encourage psychology researchers to think more thoroughly about effect sizes, recent articles have revisited the argument that effect sizes that might seem small may actually matter in our science (Funder & Ozer, 2019; Götz et al., 2021). Although similar arguments have been made in the past (e.g., Abelson, 1985), the more recent of these papers have received substantial attention, with Funder and Ozer (2019) already gaining over 863 citations, and Götz et al. (2021) 48 citations (according to Google Scholar as of February 10, 2022). Many of these citations are used to argue for the importance of research findings. For example, 31 of the 48 citations for Götz et al. (2021) are to justify the importance of some effect size. However, none of the articles citing Götz et al. (2021) consider any arguments *against* the importance of the effect size that was observed. Hence, although high-profile publications such as Götz et al. (2021) have helped move researchers away from heuristically dismissing effects that are arbitrarily classified as small, the field is now in danger of heuristically accepting all effects as important.

Götz et al. (2021; as well as Funder & Ozer, 2019) provide a sound basis for why we, as a field, should avoid classifying effects as small based on arbitrary criteria and that, to provide a foundation for a cumulative psychological science, we should instead accept such "small" effects as the norm. Our paper focuses on some of the specific arguments by Götz et al. (and by others) that seem to invoke singular mechanisms that may apply if the observed effect size were generalized to some broader real-world context outside of the study. For example, Götz et al. (2021) argue that, when considered within the dynamic contexts of everyday life and society, some effect sizes may become larger through accumulation over time or when applied to large populations (see also, Abelson, 1985; Cortina & Landis, 2009; Funder & Ozer, 2019; Ozer & Benet-Martínez, 2006; Silan, 2019). Yet it's highly unlikely that only a single mechanism would influence the effect size when the finding is generalized to new contexts. Therefore, the focus on a single mechanism that amplifies the importance of observed effect sizes is an oversimplification that can lead to excessively optimistic views about whether psychological effects matter.

Indeed, there are going to be boundary conditions for when effects will be important (Sauer & Drummond, 2020) as well as counteracting mechanisms that can make effects less important. For example, psychological reactance (Stindl et al., 2015), a type of counteracting response, may lead people to resist interventions aimed at changing their views or behaviours, which can prevent even "large" observed effect sizes from being important in the context of interest.

To be precise in language when describing different concepts, we use *observed effect size* in reference to the effect sizes estimated in some study (or studies), which we distinguish from the *generalized effect*, i.e. the expected effect if the study was done in the exact context of application with all the relevant mechanisms applying. For example, take the claim that an observed effect size is important because it would accumulate with repetition in a more dynamic context. This is essentially stating that the generalized effect will be larger than the observed effect size due to the mechanism of accumulation through repetition.

For claims about the potential importance of findings to be specific enough to be logically or empirically verifiable, researchers need to explicitly state the mechanisms that can amplify the importance of an observed effect size *and* the mechanisms that counteract it, in addition to any other relevant considerations that might influence how the effect generalizes. As such, multiple mechanisms that might operate simultaneously should be considered, and all of the relevant factors that would impact an effect's importance (i.e., auxiliary assumptions) should be stipulated so that other researchers can evaluate the claims and even test them where appropriate (Uygun-Tunç & Tunç, 2020). Researchers would thus be providing verifiable lines of reasoning to justify their claims about an effect's (un)importance.

To facilitate this process, in Table 1 we present examples of (i) mechanisms that amplify importance; (ii) mechanisms that counteract importance; and (iii) other considerations such as the target population's baseline levels on the variables of interest. Each of these mechanisms will vary in relevance across contexts and for different phenomena and will depend on the differences and similarities between the context of the study and the context to which the finding is being generalized. In the Supplemental Materials we provide more detailed descriptions of each mechanism and their relevant assumptions.

Researchers who wish to argue that an observed effect size is important, due to a generalized effect, will need to explicitly state the relevant amplifying mechanisms that will operate in the context to which they are generalizing. Moreover, they will need to explicitly state why the counteracting mechanisms are likely to *not* apply, or why the amplifying mechanisms would be stronger. Likewise, researchers who claim that an observed effect size is not important will need to communicate which counteracting mechanisms are assumed to operate and the relevant factors involved. They will further need to justify why the amplifying mechanisms will not apply, or why counteracting mechanisms will be stronger. In an ideal case, claims about the (un)importance of observed effect sizes should be based on empirical data (Primbs et al., 2021), or, at the very least, stated in a way so that the claims can be corroborated or falsified in future empirical work.

If researchers explicitly state all of the relevant mechanisms that they think would apply, then the assumptions and hypotheses behind claims about the (un)importance of effects will be made clear. Such transparent claims about effect size generalization will be more amenable to formal theoretical work, such as analytical models or simulation studies that test internal consistency and logical coherence (e.g., Smaldino, 2017). Further, communicating hypotheses and assumptions transparently will help highlight where disagreements lie and where future empirical investigations should focus. For example, a justification that explicitly rests on the assumption that an observed effect size will accumulate through repetition can be examined by studies that are specifically designed for this purpose, as well as for ruling out any counteracting mechanisms, such as habituation (Groves & Thompson, 1970). Thus, by making transparent claims about effect size generalization, evaluating the importance of research findings can become empirically or logically verifiable, perhaps even falsifiable. In the Supplemental Materials we provide a hypothetical example for how our proposal can help to move researchers away from speculation and towards empirical verification.

With increasing scrutiny placed on how we report and communicate the implications of our findings (e.g., Premachandra & Lewis, 2021), explicitly stating the assumptions and testable predictions needed to substantiate claims about effect size generalization could become recommended or required at different stages on the path from psychology to practice. For example, reviewers and editors can ask authors to justify their claims for whether an effect matters in the context to which application is being recommended. Consumers of research findings, such as policy decision makers and other stakeholders, would thus have a guide to help evaluate the quality of claims about the implications of research findings for dynamic real-world contexts.

Researchers should also keep in mind that assessing how effect sizes would generalize is not a one-time event: research findings may have implications for many different contexts that vary on relevant factors. Different contexts may thus involve different amplifying and counteracting mechanisms. Moreover, the same contexts may change with time. For example, people's attitudes and behaviours can change once they learn about psychological phenomena or with changing cultural norms (Gergen, 1973). Therefore, assessing how an observed effect size generalizes should be considered an on-going process.

Summary

Recent high-profile publications draw on existing arguments that effects typically classified as small in psychology may be important when generalized to dynamic real-world contexts where the effect size may accumulate over time or across large populations. To avoid the field heuristically accepting all effects as important, we argue that researchers' claims about the importance of effects should incorporate all of the relevant mechanisms that would amplify and counteract the observed effect size in the process of generalization. This will help make researchers aware of the assumptions necessary for their claims to hold. It will also encourage researchers to make these assumptions explicit. Claims about the importance of observed effects, and the relevant assumptions, can then become objects of investigation themselves.

Table 1. *Summary of examples to help with building verifiable arguments about a finding's practical relevance*

	Amplifying Mechanisms
Accumulation through repetition	An observed effect size can be important due to the generalized effect it has via the process of accumulation through repetition. For example, negative stereotypical views held about a person can induce that person to behave in ways that confirm those views. However, empirical studies underestimate the real impact of such self-fulfilling prophecies because the studies typically involve the effects of a single perceiver on a target person; whereas in the context of everyday life, there are likely to be multiple perceivers creating a cumulative and larger effect (Claire & Fiske, 1998). Thus, the claim is that the observed effect sizes of self-fulfilling prophecies in the lab represent larger generalized effects in the world, due to the great number of times a person or group is exposed to the phenomenon revealed by the studies.
Amplification through interaction	Sisk et al.'s (2018) meta-analysis showed that the average effect size from growth mindset interventions is "small" and unlikely to be practically important. However, a large randomized controlled trial showed that the effect of a growth mindset intervention was moderated by whether the school provided high-quality learning opportunities. The effect was therefore smaller when averaged across the whole sample and larger when examining the subsample of at-risk students from lower performing schools (Yeager et al., 2019). The effect of the intervention on at-risk students was thus amplified by whether the school was high or low achieving.
Cascades	Even if an observed effect size is small according to some arbitrary criteria, the generalized effect in a more dynamic context may involve a causal chain, or cascade, that later results in a larger effect size (i.e., downstream consequences). For example, Bond et al. (2012) ran a randomized controlled trial of political mobilization with 61 million Facebook users in the United States. Messages directly influenced people's voting behaviour, but also their political self-expression on Facebook, which cascaded into differences in the voting behaviour of other people who were close Facebook friends with the participants. Given that each participant has multiple Facebook friends, the effect size *cascaded* into a larger overall effect.
Scaling up	An observed effect size may be important because of the amount of people that the generalized effect is scaled to. For example, natural experiments examining the impact of mass media campaigns on smoking have found effects that may be classified as small according to Cohen's criteria, for example, but which are considered important due to the large number of people who subsequently stopped smoking (Durkin et al., 2012).
	Counteracting Mechanisms
Habituation	If people habituate to a stimulus, then an observed effect size from a study will be unlikely to have a generalized effect through the process of accumulating additively, and once a population is fully habituated there will be no generalized effect at all. For example, people

	become habituated to a reinforcing stimulus the more that stimulus is presented, reducing the effectiveness of the stimulus upon repetition (Lloyd et al., 2014).
Counteracting responses	People may alter their behaviour in ways to counteract the observed effect size (see also psychological reactance; Stindl et al., 2015). For example, because people want to be perceived in a way that is consistent with their own self-views, they are often motivated to influence how others perceive them so that they will sometimes respond in ways to convince perceivers holding negative stereotypical views that those stereotypes are wrong (Jussim, 2012; Swann & Ely, 1984). Therefore, observed effect sizes from studies on self-fulfilling prophecies will not always have implications for more dynamic contexts because the generalized effect involves counteracting responses.
Homeostasis	Certain aspects of a system may have a stable set-point around which there can be fluctuations but to which the system reverts over time. Hence, for an observed effect size to matter, either the system must be one that is not governed by a stable set-point or the observed effect size must be large enough to create a new set-point. For example, hedonic adaptation is where people revert to a set-point following changes in life circumstances that cause fluctuations in subjective well-being or positive and negative affect (Diener et al., 2006; Lucas, 2007). Therefore, in systems governed by homeostasis, unless the observed effect size is large enough to create a new set-point, the generalized effect may be too small and/or transient to matter.
Counteraction through interaction	Similar to amplification through interaction, there are likely to be factors in the context of application that will interact with the variables of interest to *counteract* the observed effect size so that the generalized effect is smaller (or non-existent). For example, Johnson (2012) examined the implementation of several evidence-based marriage interventions in the U.S. The results of the interventions comparing treatment groups to a control ranged from positive effects at one site where the intervention was deployed, through null effects at six other sites, to negative effects at an eighth site. The problem was that the interventions had been designed based on findings from research that used samples of middle-class couples but was then applied to poorer populations for whom marriage quality, at least as conceptualized by the interventions, was not always highly prioritized. The extent to which people prioritized marriage quality in their life may have interacted with the interventions to produce an overall negligible effect.
Other Considerations	
Baseline Levels	The target population's baseline levels on the variables of interest will determine whether an observed effect size will have a generalized effect. Milgram et al. (1969) found that the number of confederates who looked up into an empty sky had a curvilinear effect on the number of passers-by who conformed and also looked up: there was little additional impact of having more than five confederates, up to a total of fifteen.
Stage of development	If there is a developmental process involved in the phenomenon, the state of the target population needs to be accounted for to understand the generalized effect. For example, the relationship between a supervisor and employee develops over a long period. Therefore, if a study investigating supervisor behaviour and employee well-being is conducted when the supervisor and employee samples are already in their work roles then it would be missing the processes important for the development of the relationship (cf. Butler, 2011; Hofmans et al., 2019).

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

FA received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 883785. RAK was supported by a Hypatia fellowship from the RadboudUMC. DL and LT were funded by VIDI Grant 452-17-013 from the Netherlands Organisation for Scientific Research. AKP was supported by a grant provided by the Huo Family Foundation. AO and AKP received funding from the ESRC (ES/T008709/1). AO was supported by a College Research Fellowship by Emmanuel College, University of Cambridge.

ORCID iD

Farid Anvari https://orcid.org/0000-0002-5806-5654
Rogier A. Kievit https://orcid.org/0000-0003-0700-4568
Daniël Lakens https://orcid.org/0000-0002-0247-239X
Charlotte R. Pennington https://orcid.org/0000-0002-5259-642X
Andrew K. Przybylski https://orcid.org/0000-0001-5547-2185
Leo Tiokhin https://orcid.org/0000-0001-7333-0383
Brenton M. Wiernik https://orcid.org/0000-0001-9560-6336
Amy Orben https://orcid.org/0000-0002-2937-4183

Author Contributions

A. Orben and F. Anvari were part of a team who organized a hackathon at SIPS2019 in which the ideas for this paper originated; B. M. Wiernik and A. K. Przybylski attended the hackathon. F. Anvari wrote the first draft based on notes taken by B. M. Wiernik from the hackathon; F. Anvari and A. Orben revised the first and final drafts. D. Lakens, L. Tiokhin, R. Kievit, B. M. Wiernik, A.K. Przybylski, and C. R. Pennington provided critical feedback and revisions for the manuscript. All authors approved the final version of the manuscript prior to submission.

Acknowledgements

This preprint was created using the preprint template from https://osf.io/t4eqp/, created by Brenton M. Wiernik.

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin*, *97*, 129–133. https://doi.org/10.1037/0033-2909.97.1.129

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature, 489*, 295–298. https://doi.org/10.1038/nature11421

Butler, E. A. (2011). Temporal interpersonal emotion systems: The “TIES” that form relationships. *Personality and Social Psychology Review*, *15*(4), 367-393. https://doi.org/10.1177/1088868311411164

Claire, T., & Fiske, S. (1998). A systemic view of behavioral confirmation: Counterpoint to the individualist view. In C. Sedikedes, J. Schopler, & C. A. Insko (Eds.), *Intergroup cognition and intergroup behavior* (pp. 205 – 231). Mahwah, NJ: Erlbaum.

Cortina, J. M., & Landis, R. S. (2009). When small effect sizes tell a big story, and when large effect sizes don’t. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends* (pp. 287–308). New York, NY: Routledge.

Diener, E., Lucas, R. E., & Scollon, C. N. (2006). Beyond the hedonic treadmill: Revising the adaptation theory of well-being. *American Psychologist*, *61*(4), 305. https://10.1037/0003-066X.61.4.305

Durkin, S., Brennan, E., & Wakefield, M. (2012). Mass media campaigns to promote smoking cessation among adults: An integrative review. *Tobacco Control*, *21*(2), 127-138. https://doi.org/10.1136/tobaccocontrol-2011-050345

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. https://doi.org/10.1177/2515245919847202

Gergen, K. J. (1973). Social psychology as history. *Journal of personality and social psychology*, *26*(2), 309. https://doi.org/10.1037/h0034436

Götz F. M., Gosling S. D., Rentfrow P. J. (2021). Small Effects: The Indispensable Foundation for a Cumulative Psychological Science. *Perspectives on Psychological Science*. July 2021. https://doi.org/10.1177/1745691620984483

Groves, P. M., & Thompson, R. F. (1970). Habituation: a dual-process theory. *Psychological Review*, *77*(5), 419. https://doi.org/10.1037/h0029810

Johnson, M. D. (2012). Healthy marriage initiatives: On the need for empiricism in policy implementation. *American Psychologist*, *67*(4), 296. https://doi.org/10.1037/a0027743

Jussim, L. (2012). *Social perception and social reality: Why accuracy dominates bias and self-fulfilling prophecy*. OUP USA. https://doi.org/10.1093/acprof:oso/9780195366600.001.0001 . See page 252.

Lloyd, D. R., Medina, D. J., Hawk, L. W., Fosco, W. D., & Richards, J. B. (2014). Habituation of reinforcer effectiveness. *Frontiers in Integrative Neuroscience*, *7*, 107. https://doi.org/10.3389/fnint.2013.00107

Lucas, R. E. (2007). Adaptation and the set-point model of subjective well-being: Does happiness change after major life events?. *Current Directions in Psychological Science*, *16*(2), 75-79. https://doi.org/10.1111/j.1467-8721.2007.00479.x

Milgram, S., Bickman, L., & Berkowitz, L. (1969). Note on the drawing power of crowds of different size. *Journal of Personality and Social Psychology, 13,* 79–82. https://doi.org/10.1037/h0028070

Mold, J. W., & Stein, H. F. (1986). The cascade effect in the clinical care of patients. *New England Journal of Medicine, 314*(8), 512-514. https://doi.org/10.1056/nejm198602203140809

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, *57*, 401–421. https://doi.org/10.1146/annurev.psych.57.102904.190127

Premachandra, B., & Lewis, N. A., Jr. (2021). Do we report the information that is necessary to give psychology away? A scoping review of the psychological intervention literature 2000-2018. *Perspectives on Psychological Science*, online first publication. https://doi.org/10.1177/1745691620974774

Primbs, M. A., Pennington, C. R., Lakens, D., Silan, M. A. A., Lieck, D. S. N., Forscher, P. S., Buchanan, E. M., & Westwood, S. J. (2021). *There are no 'Small' or 'Large' Effects: A Reply to Götz et al. (2021)*. PsyArXiv https://doi.org/10.31234/osf.io/6s8bj

Sauer, J. D., & Drummond, A. (2020). Boundary conditions for the practical importance of small effects in long runs: A comment on Funder and Ozer (2019). *Advances in Methods and Practices in Psychological Science*, 1-3. https://doi.org/10.1177/2515245920957607

Silan, M. A. (2019). *A Primer on Practical Significance*. PsyArXiv https://doi.org/10.31234/osf.io/zdhfe

Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L., & Macnamara, B. N. (2018). To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses. *Psychological Science*, *29*(4), 549-571. https://doi.org/10.1177/0956797617739704

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational social psychology*, (pp. 311-331). Routledge, New York: NY. https://doi.org/10.4324/9781315173726-14

Steindl, C., Jonas, E., Sittenthaler, S., Traut-Mattausch, E., & Greenberg, J. (2015). Understanding psychological reactance: New developments and findings. *Zeitschrift fur Psychologie*, *223*(4), 205–214. https://doi.org/10.1027/2151-2604/a000222

Swann, W. B., & Ely, R. J. (1984). A battle of wills: Self-verification versus behavioral confirmation. *Journal of Personality and Social Psychology*, *46*(6), 1287. https://doi.org/10.1037/0022-3514.46.6.1287

Uygun Tunç, D., & Tunç, M. N. (2020). *A falsificationist treatment of auxiliary hypotheses in social and behavioral sciences: Systematic Replications Framework*. PsyArXiv. https://doi.org/10.31234/osf.io/pdm7y

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachano, R., Buontempo, J., Yang, S. M., Carvalho, C. M., Hahn, P. R., Gopalan, M., Mhatre, P., Ferguson, R., Duckworth, A. L., & Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*(7774), 364-369. https://doi.org/10.1038/s41586-019-1466-y