Character Trouble in Times of Metascientific Trouble

Felipe Romero
University of Groningen

Abstract. In this critical response to John Doris's book "*Character Trouble: Undisciplined Essays on Moral Agency and Personality*," I analyze his updated take on character skepticism—the view that character traits have surprisingly limited influence on behavior across diverse situations—from a philosophy of science perspective. While I find his updated view compelling, I challenge his reliance on Cohen's conventional effect size benchmarks, arguing that qualitative labels for effect sizes obscure rather than clarify the practical significance of results. I propose that Doris's strongest argument lies in what I call the "disproportion thesis"—the view that personality variables exert less influence, and situational variables more influence, on behavior than our intuitive expectations would predict, creating a disconcerting gap. However, I argue that this thesis requires a more explicit quantification of those prior expectations. I conclude that character skepticism would benefit from formulations of its insights in a way that directly addresses character theorists' empirical commitments, avoiding vague benchmarks and contextualizing effects.

1. Introduction

For over 20 years, John M. Doris has championed the "character skepticism" theme. Relying on psychological findings, he has explored the surprisingly limited influence of character traits on trait-relevant behavior and the import (or, better to say, lack of import) of this influence on moral philosophy. *Character Trouble* (Doris, 2022)[1] is a rich collection of essays that traces the highlights of this 20-year trajectory.[2]

The book also updates Doris's views. In the closing essay, "The Future of Character," he rearticulates character skepticism in light of a timely and careful reassessment of the empirical evidence that supports it. I find this updated version of the view largely persuasive and believe it should be compelling to those interested in discussions about character. I also find the book quite compelling from a methodological point of view. If Doris is correct that

[1] Citations containing page numbers only refer to the book.

[2] The book also explores what Doris calls "the frail agency hypothesis" (Doris, forthcoming), but my comments will be centered on character skepticism.

the merits of a philosophical method should be assessed by looking at its products (Doris, 2015, p. 12; Doris, forthcoming), his case for character skepticism is also a strong case for his methodological commitments. These commitments include a "bare-knuckle naturalism" (p. 197, p. 243), a "historical instrumentalism" (p. 195), and one that we could call *empirical provisionalism*, an appeal to continuously revise philosophical theories as new evidence appears (p. 212). Practitioners of empirically-informed philosophy may find plenty to be inspired by.

I distinguish two cases for character skepticism in "The Future of Character." I call the first the *classic* case and concerns the interpretation of specific findings supporting the view. Legitimate worries about the replicability of scientific findings in psychology and related fields have occupied methodologists since the 2010s. Hence, it is natural to wonder to what extent such worries taint empirically-informed moral psychology. Doris does a great job dispelling those worries about character skepticism with his re-reading of the classic studies, and I will not comment on them.

I call the second case *metascientific*, although Doris does not advertise it as such. With this term, I mean that the arguments appeal to considerations about the possibility of evidence and evidence about the evidence in personality and social psychology. The arguments for this case are intriguing and worth discussing. While I am sympathetic to the insights in these arguments, I will comment as a philosopher of science interested in the structure of evidence needed to support them. I will argue that the metascientific case depends at points on evidence that is, at best, implicit in the literature but likely absent from it.

2. The Metascientific Case for Character Skepticism

Effect sizes become relevant when we go beyond statistical significance as the main criterion to assess a study's import. They are also central to assessing the implications of studies outside their experimental settings: often, observed effect sizes from methodologically solid research have little practical import. This fact raises the bar for the empirically-informed philosopher. Doris is at the forefront here. He frames the discussion of character skepticism by attending to effect sizes, which constitutes a step forward.

As I see it, Doris's case has two parts. First, he proposes an "Argument from Effect Sizes," which states that "given the reasonably expectable range of effect sizes in psychology – small to moderate – character traits cannot be reasonably expected to exert a large influence on behavior." (Doris, forthcoming, p. 8).

Let's unpack this argument. Doris relies on Cohen's (1988) proposed convention to classify effect sizes qualitatively. According to this convention, in terms of the Pearson correlation coefficient, an effect of r=.1 is "small," r=.3 is "medium," and r=.5 is "large." Cohen came up with these benchmarks based on his "subjective average" (Cohen, 1988, p. 13) of effect sizes in behavioral science. The medium benchmark is chosen because, for him, many effect sizes in behavioral science are of that magnitude, and they would be "perceptible to the naked eye of a reasonably sensitive observer" (Cohen, 1988, p. 80).

Now, the Argument from Effect Sizes appeals to the observation that, since the beginning of the "person-situation" debate in psychology, it was known that the correlations between personality variables and behavioral outcomes (e.g., the "personality coefficient") were typically $r \leq .3$, that is, small to medium in Cohen's convention. In other words, personality influences such outcomes only very modestly. I will call this the *moderate influence thesis*:

> **The moderate influence thesis:**
> The effect sizes of the influence of personality variables on cross-situational behavior are, at best, small to medium.

This thesis is at odds with the character theorist's assumption that "subjects will consistently display trait-relevant behavior across a diversity of trait-relevant eliciting conditions" (Doris, 2022, p. 214). For Doris, if this assumption were true, we wouldn't systematically observe small and medium effects (as defined by Cohen's benchmarks) in the studies that assess the influence of those traits on behavioral and cognitive outcomes. In his words, "You can't squeeze robust traits out of $r \leq .3$" (p. 223).

The second part of the case strengthens character skepticism by appealing to considerations about what we can expect based on our knowledge *of* the evidence. For instance, Doris mentions a study by Kühberger, Fritz, & Scherndl (2014), which shows that the distribution of observed effect sizes in psychology is positively skewed with a mode around r=.3 and argues that the exceptions that go above this value are pretty limited in number for personality psychology. He also observes that when science's cutting edge obtains larger effect sizes, it misses the relevant issue, i.e., cross-situational behavior (pp. 225–226). Under the assumption that these observations are correct, Doris infers that the effects in personality and social psychology are not only small to medium but also likely to be, at best, small to medium in possible observations. Hence, the character theorist and the virtue ethicist

lack what's needed to support their projects, and it is doubtful they will ever have it. I will now turn to examining some issues with this argument.

3. Effect Sizes and the Naked Eye

As expressed in the moderate influence thesis, Doris heavily relies on Cohen's benchmarks to state the Argument from Effect Sizes, labeling the effect sizes $r \leq .3$ as small to medium. I will discuss this reliance. Critics have taken issue with these benchmarks, *not only with the specific cut points but with the practice itself of having general absolute classifications of effect sizes in qualitative terms*. Several methodologists have worried about the practice and advise against their indiscriminate use (Cumming, 2011, p. 384; Lakens, 2013, p. 3). Indeed, Cohen himself warned his readers and invited them not to use them if possible.[3] Thompson (2007) points out that rigid reliance on Cohen's cutoffs could lead researchers to problems similar to those caused by rigid reliance on the criterion for statistical significance. Following Lakens (2013), one could worry that authors often don't know how to interpret what they find; consequently, they use the general benchmarks in circular reasoning, e.g., the effect sizes in personality psychology are small to medium because they are $r \leq .3$, and effects sizes of $r \leq .3$ are small to medium. This use occludes plausible interpretations that depend on context: an effect size $r \leq .3$ can be large, such as an intervention that reduces suicide rates reliably with an $r = .1$ (Lakens, 2022, sec. 6.5; Myers, Well, & Jr, 2010, p. 454). Recently, some authors have encouraged researchers to justify the importance (or unimportance) of effects without relying on them (Primbs et al., 2023) or even field-specific versions of them (Panzarella, Beribisky, & Cribbie, 2021).

There is space for disagreement, and Doris is not oblivious to these issues. His extensive discussion on effect sizes includes a defense of them. While I will arrive at a similar conclusion to the critics', I will try to evaluate what I take to be Doris's case for the conventions in its terms. Doris convincingly argues that when we translate correlations $r \leq .3$ to other effect size measures, they are the same $r \leq .3$, just in a different guise (pp. 220–222).

[3] Cohen says, "To begin with, these proposed conventions were set forth throughout with much diffidence, qualifications, and invitations not to employ them if possible. The values chosen had no more reliable a basis than my own intuition. They were offered as conventions because they were needed in a research climate characterized by a neglect of attention to issues of magnitude." (Cohen, 1988, p. 532). The warning has been largely unnoticed in practice.

However, this alone has no qualitative implications. The question that needs to be answered is, why is r≤.3 small to medium? I identify two interpretations of the convention in Doris's discussion, one that he seems to reject and another that he endorses. I agree with his rejection of the first one, but I also find the second one problematic.

The first interpretation anchors the meaning of "small," "medium," and "large" in relation to a distribution of observed effect sizes. This is how many of Cohen's readers interpret his "subjective" assessment (Gignac & Szodorai, 2016; Schäfer & Schwarz, 2019). In particular, a medium effect size is one whose magnitude is a central tendency measure (e.g., the median) in the distribution of effect sizes of a research field.

It does not take much to see the problems with this interpretation because, as Doris acknowledges, one could object to the specific cut points that Cohen chose. Data from specific fields illustrate the arbitrariness: a correlation r=.3 could be large or small depending on the field. Indeed, suppose we did Cohen's intuitive exercise systematically by looking at the empirical distribution of effect sizes in personality and social psychology–the contexts Doris cares about. In that case, we may find that correlations around r=.3 are likely in the higher quantiles, meaning they would most likely be labeled as medium to large (Gignac & Szodorai, 2016; Lovakov & Agadullina, 2021). Furthermore, one metascientific consideration is that the true effect sizes should be lower than we currently observe. This can be inferred from distributions of observed effect sizes from pre-registered studies, which arguably represent the population of effects better (Schäfer & Schwarz, 2019).[4]

More substantively, this empirical interpretation of "medium" only tells us where an observed effect would be in relation to other effect sizes in the literature. This information could help researchers have an inventory of what they are finding in their fields, but it does not tell us much about what an effect itself means for the phenomenon in question.

The second interpretation is to view r=.3 as an effect "perceptible to the naked eye of a reasonably sensitive observer" (Cohen, 1988, p. 80). Doris endorses this interpretation, and it is at the cornerstone of his Argument from Effect Sizes:

> "By the time we get to r = .3, we've got a relationship, according to Cohen (1988, p. 80), that is "perceptible to the naked eye of a reasonably sensitive observer." Definitely worth knowing about. But "perceptible" isn't "destiny," or even "dramatic"; it doesn't approach the kind of

[4] The fact that the effects of personality *and* social psychology are likely smaller than what the published literature shows may lead one to think that neither personality *nor situation* should be that powerful from Doris's metascientific perspective. I leave this for another occasion.

> influence associated with traits, like the virtues, supposed to be robust. You can't wring blood from a stone, and you can't squeeze robust traits out of r≤.3." (pp. 222–223)

And he concludes some pages later:

> "[G]iven the plausibility of the "naked eye" benchmark residing around .3, Cohen's standards strike me […] as pretty reasonable." (p. 231)

My worry is this: If "the naked eye of a reasonably sensitive observer" is a good criterion to pick out correlations r=.3 (and above), then it should do so reliably across situations, even allowing for some margin of error. We cannot expect this reliability. I take this criterion to exemplify one problem with absolute benchmarks.

First, the "naked eye" is a metaphor for which Cohen does not give precise definitions but only approximates via examples:

> "A medium effect size is conceived as one large enough to be visible to the naked eye. That is, in the course of normal experience, one would become aware of an average difference in IQ between clerical and semiskilled workers or between members of professional and managerial occupational groups" (Cohen, 1988, p. 26)

In these examples, Cohen describes pretty informal and underspecified situations: the "reasonably sensitive observer" (whoever fits the bill) is out there trying to notice correlations in the wild (not in a controlled situation) and not collecting data systematically for analysis. Many factors could vary in these situations (e.g., limited samples, sample bias, outliers, recency effects, motivated reasoning–the list is long), which could mislead observers who consider themselves reasonable. However, as far as I can see, there isn't much more in Cohen to specify this metaphor further. At the bottom, there is little beyond his observational intuitions to justify the arbitrarily specific number .3.

At this point, we could ask, what can the naked eye really perceive? This is hard to answer, but we could try to find proxy measures. One such measure could come from the experimental research on the perception of correlations. In a standard paradigm, participants look at data sets presented in a scatterplot and are asked to assess whether they perceive a correlation by eye and its magnitude. This situation can be seen as a rather favorable setting for naked-eye estimation: the observer receives all the information necessary for the task in an accessible visual form. Without getting into much detail, different authors have reported multiple times that participants perceive little correlation when r<.20 and, importantly,

systematically and severely underestimate correlations, particularly when .20≤r≤.60 (Cleveland, Diaconis, & McGill, 1982; Doherty, Anderson, Angott, & Klopfer, 2007; Rensink, 2017; Strahan & Hansen, 1978).

Considering the research on the perception of correlations as relevant here requires an imaginative leap.[5] However, I am trying to approach an assessment of what the naked eye could see in the absence of anything more than the metaphor and underspecified examples. If the question is, at what point people can perceive a correlation, the answer seems to be that they already do around r=.2. One could try to argue that r=.2 is not too far from Cohen's "medium," but it is equally far from Cohen's "small." Maybe the fact that we could see some of those "small" correlations already with the naked eye doesn't make them that small. However, if accuracy matters, and it perhaps does for the character skeptic, the range of severe underestimations should be telling. Notice that r=.6 is already above what Cohen's convention would deem as "large" (i.e., r=.5). That is, at r=.6, you have a clearly "large" correlation according to the convention but still one that people would severely underestimate.[6]

If this imaginative leap is granted, it's not unreasonable to expect people to perform worse at estimating correlations in the real world. In these experiments, participants work with complete information without many practical complications. And, if people are pretty imprecise in what they can perceive with the naked eye in those conditions, what can we expect from them in real-world conditions? Moreover, one should expect even less reliability from people trying to assess correlations between, e.g., conscientiousness and behavior in the real world. The problem would arise even in the presence of true "large" correlations in Cohen's benchmark. We would likely underestimate their strength unless it is very large.

This is just one attempt at interpreting the naked eye criterion. There may be better ones. However, this one is sufficient to illustrate the point that the criterion is a vague concept in the philosophical sense. It may have clear cases and corresponding numerical values for "small" and "large" (and perhaps quite visible ones when estimating correlations based on physical quantities) but a relatively wide range of borderline "medium" cases. Hence, we

[5] For instance, one could wonder whether the mental computations involved in estimating correlations by observing data points on a scatterplot resemble what people do when inferring correlations in the real world from everyday situations (e.g., the differences in IQ from the example).

[6] Cumming (2011, pp. 382–385) shows a correlation exercise that illustrates how difficult it is to estimate a correlation in the .1 to .5 range on a scatterplot.

cannot expect its reliability. This vagueness makes it an unsuitable criterion to establish general absolute benchmarks for what counts as a small to medium effect.

Let's zoom out. Doris attempts to justify $r \leq .3$ as small to medium, as this is required for the moderate influence thesis. However, if the "medium" category ultimately depends on the naked eye criterion for his arguments, we have reasons to worry. The criterion and derived categories are questionable, even more so if they are used to ground the idea that the *importance* of many effect sizes is small to medium. These benchmarks confuse more than they illuminate the meaning of an effect or its practical significance.

4. The Disproportion Thesis

While Doris's argument from Effect Sizes relies on Cohen's benchmarks, some passages at the beginning of his discussion of effect sizes consider context. He notes that "questions of size are not absolute but comparative, and must be answered relative to some standard, or some set of expectations" (p. 217), and proceeds to articulate the insights of character skepticism in terms of comparisons to people's priors:

> To begin, we might interpret the effect sizes of personality variables in the context of people's prior probabilities, or "priors." If priors for the effect of size of personality variables are pretty high—for example, if people (tacitly or explicitly) expect that the influence of personality variables is on many occasions strikingly apparent to casual observation unaided by statistical artifice—the effects typical of personality psychology might rightly be thought rather small in comparison to the priors. [...] The crucial observation, and the one central to the character skepticism I still espouse, is that *dispositional variables have weaker influence, and situational variables stronger influence, than one should expect if one understands dispositional differences in terms of robust traits issuing in cross-situationally consistent behavior.* [Emphasis in the original]. If you like: the influence of dispositional variables falls short of, and the influence of situational variable exceeds, many people's priors." (p. 217)

In light of these passages and the ones I have already discussed, we can distinguish two theses about effect sizes that support Doris's case against the character theorist. The first one is the moderate influence thesis, stated above, which is *a thesis about the influence of personality on behavior*. The second one, which I call the disproportion thesis, is *a thesis about people's expectations about the influence of personality on behavior*. I choose the label because when talking about these expectations, Doris frequently emphasizes that there is a "disproportion" (e.g., p. 200, p. 203, p. 206, p. 209, p. 228) between influence and expectation; a

disproportion that the character skeptic finds disconcerting (e.g., the difference between what we would intuitively expect from participants in the Milgram experiment and what actually happens). In terms of effect sizes, I suggest the following working formulations. The disproportion thesis has two parts (I reiterate the moderate influence thesis for contrast):

> **Moderate influence thesis:**
> The effect sizes of the influence of personality variables on cross-situational behavior are, at best, small to medium.
>
> **Disproportion thesis:**
> (1) The effect sizes of the influence of personality variables on cross-situational behavior are disproportionately smaller than people's prior expectations about those effect sizes.
> (2) The effect sizes of the influence of situational variables on cross-situational behavior are disproportionately larger than people's prior expectations about those effect sizes.

The two theses work in tandem, but they differ in important ways. First, one can be true without the other. Second, the moderate influence thesis makes no direct claims about situational variables. Third, the disproportion thesis requires more information than the moderate influence thesis. For both theses, we can appeal to the metascientific data for the observed effect sizes. For the disproportion thesis, we need a statement of prior expectations. Specifying these priors explicitly raises some questions.[7]

First, there is a potential ambiguity: whose priors should be considered? Doris talks about "people," which I take to imply lay people's priors from these passages. However, the overall discussion also implies that a relevant standard of behavior is the *character theorist's priors*. Perhaps knowing both is useful. Perhaps the latter represent an ideal of the former. But

[7] Since the talk about priors echoes the vocabulary of Bayesian inference, it is worth mentioning that a worry here is not precisely the familiar worry about the *subjectivity of priors* in Bayesian inference. In this context, we know where the priors should come from: they should express people's beliefs about behavior. The worry is how to state them explicitly.

laypeople's priors may be wider. I think my prior expectations are closer to the situationist end.[8] The virtue ethicist's prior expectations for personality variables are likely higher.

The second issue is an informational gap. The problem doesn't go away by rephrasing the theses using Cohen's conventions, i.e., saying that the people's prior expectations are "large" for (1) and "small" for (2). First, if I am right in the previous section, those labels don't make much sense independently of specific experiments. Second, and more pressing in this context, if we follow Doris in the metascientific remark that the observed effects for personality and social psychology are at best small to medium, it's difficult to hold that the observed effects of the influence of situational variables are disproportionately larger than "small" but still less than "medium."

To assess the strength of the disproportion, it is necessary to have a quantitative sense of the priors, even if it's approximate. I interpret the disproportion to mean that for (1) and (2), *the difference between observed effect size and prior effect size should itself be a "large" effect size in its context*. Otherwise, the disproportion wouldn't be disconcerting. We may have intuitions for specific cases about the priors for (2). For instance, in the Milgram experiment, most people (including Milgram himself) likely expect that few participants (if any) would deliver the maximum shock level. However, as I understand Doris's presentation of the evidence, we don't have explicit quantitative information on the priors, particularly to assess (1).

Lastly, in line with the previous section, it would help to see the expected disproportion in a way that makes it clearly disconcerting. The evidence that supports (1) and the evidence that supports (2) are not equally helpful in this respect. The former is expressed in terms of Pearson correlation coefficient and Cohen's benchmarks, which, as I have discussed, require interpretation in context. The case for (2) fares better, at least in the case of Milgram. There are observed effect sizes reported as the percentage of participants who deliver deadly electroshocks, a measure much easier to interpret and arguably suitable for empirical elicitation of prior expectations. Perhaps a measure more suitable for the character skeptic would have to be a measure of the disproportion itself, that is, *a measure of how many*

[8] I think this would have been my intuition before learning anything about moral psychology, but a situationist might believe that having been a teaching assistant for Doris's courses at some point during my doctoral training and reading his work now could have also influenced my priors, making my intuitions unsuitable for prior elicitation.

people in the experiment behaved in a way consistent with the expectation, such as the one by Grice et al. (2020).

Having raised all these worries, I still find that the most persuasive case for character skepticism would appeal to the disproportion thesis (or some refined version of the working formulation I sketched). Unlike the moderate influence thesis, the disproportion thesis highlights that something is surprising in the classic studies that support character skepticism, and its case would be strengthened by making explicit the prior expectations that lead to the surprise. The gap in information about priors is understandable. Most of the evidence in personality and social psychology has been generated without quantifying priors. Hence, this is not so much an objection as an invitation to fill the gap. The exercise that Doris proposes strikes me as epistemically on the right track. The worry is that it may require more evidence than currently available or at least a refined quantitative expression of the available evidence.

5. Closing Remarks

I'll close by stating my comments in a slightly different way. Closer attention to testability conditions would strengthen the case for character skepticism. Doris argues that quantitatively and empirically, character theory is indefensible. This raises the question of what would be quantitatively and empirically necessary to support character theory. Doris's reading of the metascientific data on effect sizes in psychology, in conjunction with Cohen's benchmarks, suggests that the typical effect sizes of psychology will not be enough. However, relying on general benchmarks to argue whether an effect size is not enough (or enough) is problematic. The existing evidence informs the question of how large the effect sizes of the influence of personality/situation on behavior are. However, arguments based on the structure of this evidence address character theory only indirectly. A direct evaluation requires testing how different that influence is from the character theorist hypothesized influences. But what are these hypotheses? Notice that the difference is not simply one of framing. Since we are focusing on effect sizes, we need hypotheses about the character theorist's alleged empirical commitments. These hypotheses shouldn't rely on vague benchmarks and should allow us to make sense of specific (non) effects in their contexts. At that point, we would see how empirically unsound character theory is and whether it has been convincingly refuted.

References

Cleveland, W. S., Diaconis, P., & McGill, R. (1982). Variables on scatterplots look more highly correlated when the scales are increased. *Science*, *216*(4550), 1138–1141. https://doi.org/10.1126/science.216.4550.1138

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

Cumming, G. (2011). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge. https://doi.org/10.4324/9780203807002

Doherty, M. E., Anderson, R. B., Angott, A. M., & Klopfer, D. S. (2007). The perception of scatterplots. *Perception & Psychophysics*, *69*(7), 1261–1272. https://doi.org/10.3758/BF03193961

Doris, J. M. (2015). *Talking to our selves: Reflection, ignorance, and agency*. Oxford: Oxford University Press.

Doris, J. M. (2022). *Character Trouble: Undisciplined Essays on Moral Agency and Personality*. Oxford, New York: Oxford University Press.

Doris, J. M. (forthcoming). Precis of Character Trouble. *Philosophia*. Forthcoming.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Grice, J. W., Medellin, E., Jones, I., Horvath, S., McDaniel, H., O'lansen, C., & Baker, M. (2020). Persons as Effect Sizes. *Advances in Methods and Practices in Psychological Science*, *3*(4), 443–455. https://doi.org/10.1177/2515245920922982

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size. *PLOS ONE*, *9*(9), e105825. https://doi.org/10.1371/journal.pone.0105825

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

Lakens, D. (2022). *Improving Your Statistical Inferences*. Retrieved from https://lakens.github.io/statistical_inferences/.

Lovakov, A., & Agadullina, E. R. (2021). Empirically derived guidelines for effect size interpretation in social psychology. *European Journal of Social Psychology*, *51*(3), 485–504. https://doi.org/10.1002/ejsp.2752

Myers, J. L., Well, A. D., & Jr, R. F. L. (2010). *Research Design and Statistical Analysis: Third Edition* (3rd ed.). New York: Routledge. https://doi.org/10.4324/9780203726631

Panzarella, E., Beribisky, N., & Cribbie, R. A. (2021). Denouncing the use of field-specific effect size distributions to inform magnitude. *PeerJ*, *9*, e11383. https://doi.org/10.7717/peerj.11383

Primbs, M. A., Pennington, C. R., Lakens, D., Silan, M. A. A., Lieck, D. S. N., Forscher, P. S., … Westwood, S. J. (2023). Are Small Effects the Indispensable Foundation for a Cumulative Psychological Science? A Reply to Götz et al. (2022). *Perspectives on Psychological Science*, *18*(2), 508–512. https://doi.org/10.1177/17456916221100420

Rensink, R. A. (2017). The nature of correlation perception in scatterplots. *Psychonomic Bulletin & Review*, *24*(3), 776–797. https://doi.org/10.3758/s13423-016-1174-7

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*. https://doi.org/10.3389/fpsyg.2019.00813

Strahan, R. F., & Hansen, C. J. (1978). Underestimating Correlation from Scatterplots. *Applied Psychological Measurement*, *2*(4), 543–550. https://doi.org/10.1177/014662167800200409

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, *44*(5), 423–432. https://doi.org/10.1002/pits.20234