



Barrett View #6

The reporting of effect sizes: Will those who understand nothing about explanatory accuracy please remain silent

The paper I report upon below reminded me of the title of David Lykken's insightful chapter entitled: [What's Wrong with Psychology Anyway?](#) [Lykken, D.T. (1991). In D. Cicchetti and W.M. Grove. (Eds.). *Thinking Clearly about Psychology. Volume 1: Matters of Public Interest. University of Minnesota Press*], except I'd now rewrite the title as: [What's Wrong with Psychologists Anyway.](#)

A colleague brought this ghastly article to my attention:

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, 2, 156-168.

Abstract

Effect sizes are underappreciated and often misinterpreted—the most common mistakes being to describe them in ways that are uninformative (e.g., using arbitrary standards) or misleading (e.g., squaring effect-size rs). We propose that effect sizes can be usefully evaluated by comparing them with well-understood benchmarks or by considering them in terms of concrete consequences. In that light, we conclude that when reliably estimated (a critical consideration), an effect-size r of .05 indicates an effect that is very small for the explanation of single events but potentially consequential in the not-very-long run, an effect-size r of .10 indicates an effect that is still small at the level of single events but potentially more ultimately consequential, an effect-size r of .20 indicates a medium effect that is of some explanatory and practical use even in the short run and therefore even more important, and an effect-size r of .30 indicates a large effect that is potentially powerful in both the short and the long run. A very large effect size ($r = .40$ or greater) in the context of psychological research is likely to be a gross overestimate that will rarely be found in a large sample or in a replication. Our goal is to help advance the treatment of effect sizes so that rather than being numbers that are ignored, reported without interpretation, or interpreted superficially or incorrectly, they become aspects of research reports that can better inform the application and theoretical development of psychological research

Except where they at least talk some sense (e.g. p.165, under the heading: **Report effect sizes in terms that are meaningful in context**), all we have is the usual promotional handwaving and old-chestnut examples rolled out showing how “cumulative” tiny effects can be important in a “population”.

Note, not one scatterplot of what a correlation of 0.05 or even 0.20 actually looks like to a decision-maker, what an effect size amounts to as ‘evidence’ for accurate causal explanatory theory, or even an example of a “typical” psychological result computationally evolved over time (via simulation) to show the hypothesised cumulative consequences of an effect in a population (*which is challenging to even conceive of doing with most one-shot psychology results*).

And not a single mention of any of Chris Ferguson’s publications and examples in this area, not least the one massively contraindicating the results of Meyer et al, 2001 (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.)

And no mention of Tryon, W.W., Patelis, T., Chajewski, M., & Lewis, C. (2017). [Theory construction and data analysis](#). *Theory and Psychology*, 27, 1, 126-134, which discusses in detail the relevance of small effects in what they call a ‘web of causation’ ... but which requires a fundamentally different approach to conducting psychological research.

Or Lamiell, J.T. (2009). [The characterization of persons: some fundamental conceptual issues](#) (Chapter 5, pp 72-86). In P.J. Corr & G. Matthews (Eds.). *Personality Psychology*. New York: Cambridge University Press; and Lamiell’s exposition of the fundamental flaw in Epstein’s pronouncements on statistical aggregation and persons.

This Funder & Ozer article is just the same old stuff we see published in this area by the usual clutch of status-quo psychologists; in contrast to Ferguson’s clear-headed criticism of this nonsense [in Ferguson, C.J. (2015). ["Everybody knows psychology is not a real science": Public perceptions of psychology and how we can improve our relationship with policymakers, the scientific community, and the general public](#). *American Psychologist*, 70, 6, 527-542].

All of which is why I wrote my earlier Barrett-View on these matters a year or so ago:

[The accurate reporting of small effect sizes: A matter of scientific integrity](#)

and showed graphically (and computationally) the actual “model” [explanatory consequences of tiny r-squares](#) (deviation r-squares but r-squares nevertheless).

The real problem is that this Funder and Ozer article will be used by many as a justification for reporting their own statistical trivia.

I can foresee attempts by some to justify results possessing low explanatory/low accuracy value by invoking that claim of “epidemiological/cumulative” consequences; failing to recognize that such a justification requires empirical evidence **for each specific effect**; showing that the claimed importance for the effect is actually plausible when modelled appropriately with an evolved-over-time computational model. Without those empirical hard yards, or empirical-actuarial epidemiological evidence, all we have is subjective handwaving, appeals to ‘authorities’ like Funder & Ozer, and wishful thinking.

And this bit on page 166, in their recommendations section, entitled: **Stop using empty terminology** is just weird:

“It is far past time for psychologists to stop squaring *r*s so they can belittle the seemingly small percentage of variance explained and to stop mindlessly using J. Cohen’s (1977, 1988) guidelines, which even Cohen came to disavow. Ideally, words such as small and large would be expunged from the vocabulary of effect sizes entirely, because they are subjective and often arbitrary labels that add no information to results that can be reported quantitatively. This goal is probably unrealistic; indeed, in this article we have been unable to avoid the liberal use of these descriptive adjectives ourselves. But at the very least, it would be good to become in the habit of responding to characterizations of effect sizes as being small or large with questions about the implied comparison: The effects are small or large compared with what? Compared with what is usually found, with what other studies have shown, or with what it is useful to know? Or is another standard altogether being used? Whatever the standard of evaluation is, there ought to be one”

That question “The effects are small or large compared with what?” is irrelevant, except when it might be useful in some context as a secondary source of information.

Why, because if one chooses to use this kind of statistical parameter as an indicator of “effect”, it can be directly interpreted in terms of how well the phenomena under examination are accurately described by whatever analysis/model has been conducted/applied to data.

The wise among us and perhaps those with a focus on decision-makers needing reputable information upon which to base critical decisions, go one step further; usually by some form of

actuarial analysis or “observation-oriented” exposition, to show the consequences of that effect in the metric/event outcomes characterised by the data.

What we never do is make statements of the form: “[compared to the bulk of effect sizes found in psychological research, our effect is OK \(or whatever other euphemism is used to describe the effect magnitude\)](#)”.

What actually matters is the explanatory accuracy of the theory-claims and/or phenomena. The fact that much of psychological research is hopelessly inaccurate for explaining any consequential outcome is hardly a reason to use these frequently occurring small effect-sizes as relative standards of “importance.”

If we truly believe a tiny/small effect is important, then we must do the hard computational modeling yards to show it is indeed plausible to make such a claim, or, acquire some empirical actuarial evidence that demonstrates the validity of our claim.

Otherwise we are no different to religious ‘authorities’ making pronouncements based upon what amounts to the “laying on of hands” by “Credentialed Persons” (as Paul Meehl noted many years ago: Meehl, P. (1997). [Credentialed persons, credentialed knowledge](#). *Clinical Psychology: Science and Practice*, 4, 2, 91-98).

And, for those who keep pointing to the old Taylor-Russell tables (as do Funder & Ozer): Taylor, H.C., & Russell, J.T. (1939) [The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables](#). *Journal of Applied Psychology*, 23, 5, 565-578, they need to read the 2010 Deloitte study-report: “[A Random search for excellence: Why “Great Company” research delivers fables and not facts](#)”. That study put paid to all the simple-minded ‘feel-good’ claims about the success of selection psychometrics. As you’d expect, this study was ignored by the majority of I/O psychologists who could not seem to grasp what this work revealed about the actual importance of psychometric selection procedures to *enduring* organizational success, and the lack of real-world utility of the Taylor-Russell tables.

Reading a recent (August, 2019) [New Scientist article](#) on the comprehensive rethinking/reworking of quantum theory by theoretical physicists, I was struck by the contrast with the Funder and Ozer paper which keeps repeating the “same-old” arguments as though these are coherent answers to how we should be trying to deal with theory-claims/tests, explanations, and “magnitude-of-effect” interpretations in psychological research.

Indeed, that Tryon et al article I quoted above might just hold the key to how we begin to rethink our entire ontological approach to generating causal explanatory theory in the domain of human psychology.

Ah well, I know most will simply shrug and move on ... but the purposeful ignorance of facts and knowledge always comes with a price to be paid at some point in time, as we now have found out with the current replication crisis.