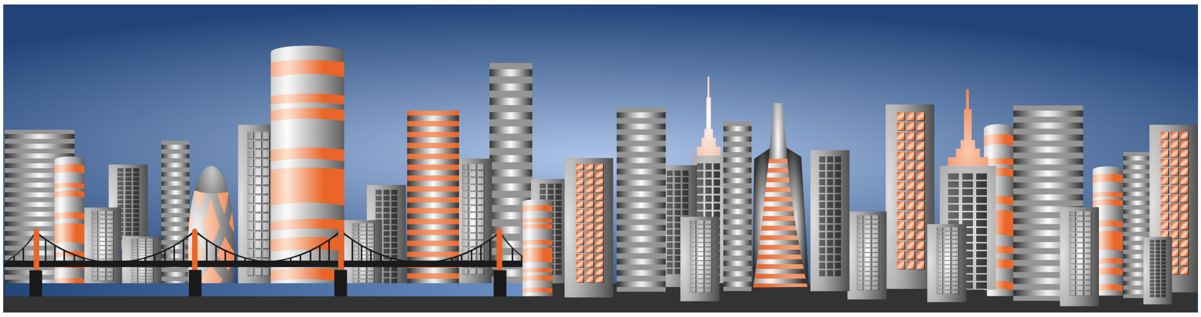

Barrett View #4

The accurate reporting of small effect sizes: A matter of scientific integrity

Many psychologists seem to have a blind-spot when it comes to interpreting their effect sizes qualitatively; usually resulting in tiny/trivial effect sizes being interpreted as near-deterministic effects, or used as evidence in support of theory-claims or statements of claim concerning prediction of some phenomenal outcome. This is not acceptable; it is a matter of scientific integrity, a matter of being truly honest and careful in the meaning/importance to be attributed to a particular effect.

Small effects are by definition an indication of inaccuracy of explanation/prediction of a phenomenal outcome. Period.

There is no point dressing them up as 'significant' when they can never be so - even at an epidemiological level, unless the phenomenon being predicted is of such criticality that any 'above chance' incidence is worth protecting against.

What brought this to a head for me was a 2016 publication in Personality and Individual Differences - partly because of what I've said *(and showed empirically)* about trivial effect sizes in my recent technical whitepaper (June, 2016) entitled: Hierarchical Multiple Linear Regression and the correct interpretation of the magnitude of a Deviation R-square.

The paper in question is Gignac, G., & Szodorai, E.T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, 102, 74-78.
Abstract
Individual differences researchers very commonly report Pearson correlations between their variables of interest. Cohen (1988) provided guidelines for the purposes of interpreting the magnitude of a correlation, as well as estimating power. Specifically, r=0.10, r=0.30, and r=0.50 were recommended to be considered small, medium, and large in magnitude, respectively. However, Cohen's effect size guidelines were based principally upon an essentially qualitative impression, rather

than a systematic, quantitative analysis of data. Consequently, the purpose of this investigation was to develop a large sample of previously published meta-analytically derived correlations which would allow for an evaluation of Cohen's guidelines from an empirical perspective. Based on 708 meta-analytically derived correlations, the 25th, 50th, and 75th percentiles corresponded to correlations of 0.11, 0.19, and 0.29, respectively. Based on the results, it is suggested that Cohen's correlation guidelines are too exigent, as <3% of correlations in the literature were found to be as large as r = 0.50. Consequently, in the absence of any other information, individual differences researchers are recommended to consider correlations of 0.10, 0.20, and 0.30 as relatively small, typical, and relatively large, in the context of a power analysis, as well as the interpretation of statistical results from a normative perspective.

Interestingly, no reference/mention at all in this more recent article to the recommendations, arguments, empirical evidence, and reasoned discussion in:

Ferguson, C.J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, 40, 5, 532-538.

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.

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.

Or even: Bosco, F.A., Aguinis, H., Singh, K., Field, J.G., & Pierce, C.A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, 100, 2, 431-449.

And yes, we have the usual "small effect sizes can be important as indicators of a phenomenal cause/relationship" line of argument. buttressed by a couple of the usual articles rolled out like old soldiers on an annual parade.

What do these articles actually indicate?

I looked more closely at the justification references for small effects articles quoted in the Gignac and Szodorai article:

Noftle, E. E., & Robins, R.W. (2007). Personality predictors of academic outcomes: Big Five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology*, 93(1), 116–130.

My attention was caught by:
"Table 9 shows the results of multiple regression analyses predicting college GPA, in

which gender, high school GPA, and SAT scores were entered at Step 1 and the Big Five dimensions were entered at Step 2. Adding the Big Five dimensions at Step 2 produced a significant increase in R-squared in all three samples." p.125, para 2

and

"Second, small effects are to be expected when predicting a multiply determined outcome (Ahadi & Diener, 1989), and academic achievement is a quintessential example of such an outcome. In our own data, we saw that when personality and SAT test scores are combined to predict college GPA, the predictive validity can reach moderate to high levels." p. 127, last para, 2nd column From Table 9 in the article, the Step 1 and Step 2 R-squares are:

	Step 1 R-square	**Step 2 R-square**	**Deviation R-square**
Sample 1	.13	.16	**.03**
Sample 2	.31	.36	**.05**
Sample 3	.16	.21	**.05**

So, as we now know from my whitepaper, such deviation r-squares convey nothing of importance, when one computes the predictive values in the metric of the actual observations, with and without the Step 2 variables *(as a side-note, the R-squares actually need adjusting for the number of predictors in the models but who cares - they are already tiny; nothing is served by making them even smaller)*.

Then we come to the second article:

Ozer, D. J., & Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, 57, 401–421.

Amazingly not one single quantitative effect size is reported in this review, just - and + signs as indicators of effect!

Are we scientists or call-center insurance salespeople? Can we not write as scientists rather than journalists, even in our annual reviews?

Quoting this article is like quoting the claims of the leadership team at Enron Inc when looking for evidence of the benefits of financial prudence.

Both articles are actually concerned with epidemiological effects, but without carefully examining the cost-benefit of the effect in a population both end up hand-waving and making vague generalizations; unlike the famous Aspirin study with an effect size of **0.52** (not 0.03) and where the outcome was mortality. That more correctly computed effect size from Ferguson (2009) puts what we actually now know about aspirin’s benefits *(and causative processes)* into a proper context.

These kinds of articles by Hemphill (2003), Bosco et al (2015), and now Gignac and Szodorai, reveal just how far some academics have drifted away from any concerns about explanatory accuracy/meaningfulness into a world dominated by the mere clerical aggregation of numbers.

To repeat what I said above, small effects are by definition an indication of inaccuracy of explanation/prediction of a phenomenal outcome. Period. There is no point dressing them up as ‘significant’ when they can never be so - even at an epidemiological level, unless the phenomenon being predicted is of such criticality that any ‘above chance’ incidence is worth protecting against.

The statements by Gignac and Szodorai on p. 76, para 2, column 2, are also of interest - for what they convey about the distinction between fact and speculation:

"For example, Ozer and Benet-Martinez (2006) argued that, even though the correlation between agreeableness and volunteerism is small (r~0.20), a slight upward shift in agreeableness at the population level *may* imply an increase of 1000s of volunteers for various organizations. Additionally, Noftle and Robins (2007) contended that the relatively small association between conscientiousness and academic achievement *may*, nonetheless, result in meaningful practical differences in the lives of individuals low and high on conscientiousness across a lifetime, based on the notion of cumulative continuity."

Anybody can say "this *may* happen". But the scientists among us take care to demonstrate it actually can happen by developing a computational model which parameterizes their thinking in a set of initial conditions and interacting functional relations, and then evolves the model over time, showing that the phenomenal outcomes they say 'may' happen over a lifetime do actually occur as expected (or not).

Failing that, you compute *(as I do)* the consequences and meaning of inaccuracy at the level of your observations, not some parameterized version of them.

In short, you do not make claims about tiny effects being important unless you can show clear empirical evidence and consequential outcome analysis to back up those claims.

posted 27th January, 2017