Supplementary Document

Gelman and Carlin (2014) argue that "when researchers use small samples and noisy measurements to study small [underlying] effects . . . a significant result is often surprisingly likely to be in the wrong direction and to greatly overestimate an [underlying] effect" (p. 1). As a demonstration of their point, suppose in a pre-post design with 20 participants (*small* as required by Gelman and Carlin [2014]), the effect of an L2 treatment on improving the use of type III conditionals (see Shintani, Ellis, & Suzuki, 2014) in English is investigated. For the sake of demonstration, let us assume that the underlying effect of this treatment in the actual population of the target learners is quantified by a standardized mean difference effect size (i.e., Cohen's *d*) of **.1** (*small* as required by Gelman and Carlin [2014]). For any statistically significant result (at .05 significance level) that we obtain from such a pre-post design, one of the following two things has likely happened. We either (a) have overestimated the underlying effect of **.1**, or (b) have found an effect that in addition to being exaggerated, is in the opposite direction to the underlying effect. Figure 1, shows how each of these two errors could happen under two scenarios.

First, suppose we obtain a statistically effect size of d = **.6** (t (19) = 2.6832, p = .0147) from our pre-post design. Figure 1 (top-panel), shows the distribution of the effect sizes for our study under H_0 (before knowing the size of the underlying effect) which assumes that the underlying effect size is "0". The position of the obtained effect size **.6** shown as a green rhombus in the right red-shaded tail. Based on this distribution, any effect size value larger than |**.468**| is called statistically significant. Yet in reality we assumed (known for the sake of demonstration) that the underlying effect size is **.1**, and thus the actual distribution of effect sizes has to look like the bottom-panel NOT the top-panel H_0. This shifting of distribution from H_0 to a distribution that assumes an underlying effect size other than "0" (here that other value is **.1**) is best known as

"***Power Analysis***" in research. If we did not assume that the underlying effect is of some size (here **.1**), we would not be realizing that we have an error of magnitude or sign in our estimation process. With this *assumed* knowledge, we translate *everything* (including the two shaded regions) from our distribution of effect size under H_0 to that under the assumed underlying effect size. Now, we have the two formerly shaded regions but this time under our new distribution under the underlying effect size. Note that under this assumed distribution the shaded regions have disproportionate sizes because the bottom-panel distribution has shifted to the right. According to this assumed distribution, on average, we will be exaggerating the magnitude of the underlying effect by a factor of 5.8 whenever our obtained results are statistically significant at .05 significance level. The calculations are found at (https://github.com/izeh/l/blob/master/9.r).

Second, suppose we had obtained the same statistically significant result but in a different direction i.e., $d = -$ **.6** (t (19) = – 2.6832, p = .0147) from our pre-post design. This time, the obtained effect is positioned in the red-shaded area in the left tail of the assumed distribution (the bottom-panel). This statistically significant result, however, in addition to overestimating the actual size of the underlying is in the wrong direction. Given the assumed underlying distribution of effect sizes, it is possible to determine the likelihood that any one significant result is in the wrong direction. In our case, out of all possible statistically significant findings that could be obtained, the likelihood that any one significant result is in the wrong direction is simply the area of the red-shaded area in the left tail divided by total areas of significance under the assumed underlying distribution of effect sizes. In this case, this likelihood is 12%.

Therefore, on average a statistically significant result from this study due mainly to its small sample size (20 participants) can either be an exaggeration of the underlying population effect by

a factor of 5.8 (i.e., 5.8 times larger than the actual underlying effect) and in addition to that there is 12% probability for that statistically significant result to be in the wrong direction as well.

Gelman and Carlin (2014) call the factor by which one can obtain a statistically significant effect that overestimates the underlying effect a "*Type M*" error (*M* for magnitude), and the probability by which one can obtain a statistically significant effect that in addition to being a *Type M* error is in the wrong direction to the underlying effect a "*Type S*" (*S* for sign) error. An R function to graphically compute all these errors and draw Figure 1 is publicly available at (https://github.com/izeh/l/blob/master/10.r).

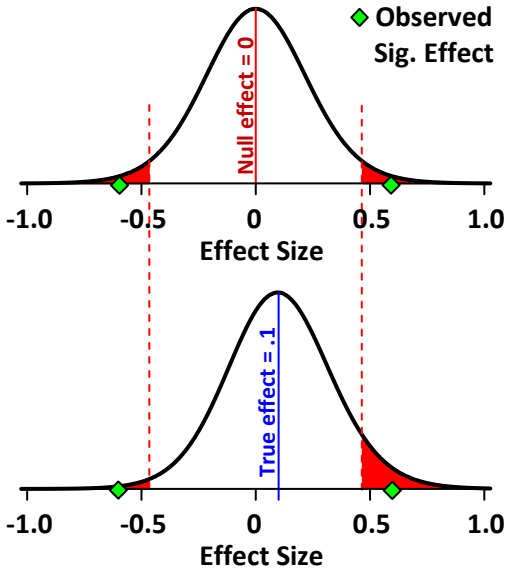

Figure 1. Type *"M"* and Type *"S"* error

Note that the proposal by Gelman and Carlin (2014) could also serve to remind us that if we detect a statistically significant finding using say, .005, as our significance level, then the idea that the underlying effect is small could not quite hold. In other words, for a finding that is significant at such a stringent significance level (i.e., .005), one is more likely to think that there is an underlying effect that is not as quite small as initially assumed. This in and by itself is

another advantage for lowering the threshold for detecting statistically significant results as proposed by several recent studies (Benjamin et al., in press; Johnson, 2013).

References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., . . . Johnson, V. E. (in press). Redefine Statistical Significance. *Nature Human Behavior*.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science, 9*(6), 641-651.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences, 110*(48), 19313-19317. doi:10.1073/pnas.1313476110

Shintani, N., Ellis, R., & Suzuki, W. (2014). Effects of written feedback and revision on learners' accuracy in using two English grammatical structures. *Language Learning, 64*(1), 103-131.