PERSPECTIVE
The New Statistics for Neuroscience Majors: Thinking in Effect Sizes

Robert J Calin-Jageman
Psychology Department, Dominican University, River Forest, IL 60305.

An ongoing reform in statistical practice is to report and interpret effect sizes. This paper provides a short tutorial on effect sizes and some tips on how to help your students think in terms of effect sizes when analyzing data. An effect size is just a quantitative answer to a research question. Effect sizes should always be accompanied by a confidence interval or some other means of expressing uncertainty in generalizing from the sample to the population. Effect sizes are best interpreted in raw scores, but can also be expressed in standardized terms; several popular standardized effect score measures are explained and compared. Reporting and interpreting effect sizes has several benefits: it focuses on the practical significance of your findings, helps make clear the remaining uncertainty in your findings, fosters better planning for subsequent experiments, fosters meta-analytic thinking, and can help focus efforts on protocol optimization. You can help your students start to think in effect sizes by giving them tools to visualize and translate between different effect size measures, and by tasking them to build a 'library' of effect sizes in a research field of interest.

Key words: inferential statistics, neuroscience education, effect sizes, confidence intervals

This is part two of a Stats Perspectives Series for JUNE.

Statistically speaking, the times they are a changing. Fueled in part by concerns over the rigor and reproducibility of the neuroscience literature, there are major changes afoot in the norms for statistical analysis. My goal in this series is to describe some of these changes and illustrate how you can help prepare your students for the new normal.

One of the most prominent trends in statistical reform is a new emphasis on reporting and interpreting **effect sizes**. For example, last year the *Journal of Neuroscience* announced updated guidelines for authors that require reporting "complete results of the statistical analyses, including… effect sizes" (Picciotto, 2017). Similarly, the American Psychological Association introduced a new emphasis on reporting effect sizes and confidence intervals in the 6th edition of its publication manual (2010; these changes are summarized in Fidler, 2010). This was reinforced in the APA's newly updated standards for reporting quantitative research (Appelbaum et al., 2018), which enjoin authors to report "effect size estimates and confidence intervals."

What is an effect size?

An effect size is very simple: it is just the quantitative answer to your research question (Box 1).

Box 1: Effect sizes

- An effect size provides a quantitative answer to your research question. Any p value has a corresponding effect size and confidence interval.
- Always report effect sizes with a confidence interval or some other way of expressing uncertainty about generalizing from the sample to the population.
- Effect sizes can be standardized but are usually best understood expressed in the same units as the dependent variable.

Let's consider an example. Dulawa et al. (2004) investigated the effect of Prozac in a mouse model of depression. BALB/c mice were treated with 10mg/kg Prozac (n=13) or placebo (n = 13). Then both groups were tested in the Porsolt swim test. Dulawa et al. (2004) reported that there was a significant decline in immobility in the group treated with Prozac ($t(24) = 2.06$, $p = 0.05$; this is based on data extracted from their Figure 2B). This is the traditional way of reporting the result. It gives us a *qualitative* research conclusion*:* it is likely that Prozac affects behavior in the Porsolt swim test. The obvious follow-up question should be: *by how much*? That's where the effect size comes in—it provides a quantitative answer to the research question. In this case, placebo-treated mice were immobile for 134s of the 240s test (s = 30). Prozac-treated mice were immobile for 94s of the 240s test (s = 58). The effect size, in this case, is the difference between these means:

$$M_{Prozac} - M_{\text{Placebo}} = \text{-40s, 95\% CI[-80, -0.008]}$$

Note that an effect size is reported with a *point estimate* and with a *confidence interval.* The point estimate is the effect found in this particular sample, in this case a 40s reduction in swim immobility. The confidence interval, on the other hand, expresses the uncertainty in generalizing from this sample to the population at large. In this case, the confidence interval is very long, reflecting considerable uncertainty. If the true effect of Prozac was an 80s reduction in immobility, these data would not be especially surprising. Similarly, these data would not be very surprising if the true effect of Prozac was vanishingly small, just a reduction of 8 thousands of a second.

The effect size and confidence interval give us valuable additional context for understanding the result reported by Dulawa et al. (2004). Whereas, the p value tells us about the statistical significance (the data are unlikely given a null hypothesis of exactly 0 effect), the effect size and confidence interval focus our minds on the *practical significance* of the finding, which in this case is highly uncertain: Prozac could have anywhere from a large to a vanishingly small impact in this model of depression. Seeing the effect size and confidence interval, we would

Box 2: Common statistical tests and their corresponding effect sizes

In the examples below the traditional hypothesis test has been replaced with an effect size (in bold) and confidence interval.

- **T-tests.** The effect size for a t-test is the difference between the two means. Example from Conte et al. (2017): *Reflexes responses were 8.9s before training and 16.5s after training, an average increase of **7.5s** 95% CI [6.8, 8.2].*
- **2x2 Interaction in a Factorial ANOVA**. The effect size is the 'difference of the difference,' or the comparison of the two simple effects. Example from Perez et al. (2018): *In naive animals, the weak shock produced a 3% decrease in reflex responsiveness. In previously trained animals, the weak shock produced a 23% increase in reflex responsiveness. Thus, there was a strong interaction between shock and previous training (**$M_{\Delta\Delta}$ = 25%** 95% CI [18,32]).*
- **2x2 Chi Square Test.** The effect size is the difference in proportions. Example modelled after Weissman et al. (1996): *In our sample, 35 of the 500 women were classified as having clinical depression (P_{women} = .07), whereas only 14 of 500 men had the same status (P_{men} = 0.028). Thus, there was a higher prevalence of depression among women: **$P_{women-men}$ = 0.04** 95% CI[.015,.07].*

want a larger sample or replication before trusting that there is a meaningful effect of Prozac on swim immobility in this strain of mouse.

Behind every traditional hypothesis test and *p* value lurks an effect size and confidence interval, just waiting to be reported and interpreted. Box 2 presents some common inferential statistics and the corresponding effects sizes. In interpreting an effect size, it is important to remember that its confidence interval represents the *mathematically ideal* range of expected sampling error. Given that real-world experiments rarely live up to the mathematical ideal (in terms of random sampling, perfect measurement, perfectly normally distributed data, etc.), we should treat the confidence interval as optimistic (see, for example, Shirani-Mehr et al., 2016).

Effect sizes can be expressed in standardized units

In general, it is best to express and interpret effect sizes in 'raw scores'—in the same units as the dependent variable. This is generally best because scientists familiar with the assay will have a good sense of the measurement scale and can best judge the practical significance of the effect size and its confidence interval.

With that said, it can sometimes be helpful to express effect sizes in standardized units. Figure 1 graphically compares some popular standardized effect size measures, each of which is explained below.

Cohen's d expresses the difference between two groups in standard deviation units. For both between- and within-subjects designs Cohen's d is calculated as:

$$d = (Mean_{Group1} - Mean_{Group2}) / s_{pooled}$$

where s_{pooled} is the standard deviation pooled across the two groups.

Cohen's d from a sample is slightly up-ward biased relative to the true population effect size. Therefore, there is a slight correction usually applied to Cohen's d. Confusingly, some researchers then label the bias-adjusted value Hedges *g*. To avoid this confusing situation, a better label is $d_{unbiased}$ (Cumming and Calin-Jageman, 2017).

In the Dulawa et al. (2004) experiment, the effect of Prozac on swim immobility is: $d_{unbiased}$ = -0.78 95%CI [-1.6, -0.00]. This means that in the sample, the reduction in swim mobility in the Prozac group was 0.78 standard deviations, a difference typically considered quite large. Again, though, the confidence interval is very wide, indicating tremendous uncertainty about the true magnitude of the effect. In general, interpretation of the raw score confidence interval and a standardized confidence interval should lead to similar conclusions—after all, it is the same information just expressed in different units.

r^2 expresses the proportion of variance shared or accounted for in the *linear* relationship between two variables; values range between 0 (fully independent) and 1 (fully collinear).

In the Dulawa et al. (2004) experiment, the effect of Prozac on swim immobility is: r^2 = 0.13 [0.00, 0.39]. This means it is plausible that Prozac explains anywhere from 0 to almost 40% of variation in swim immobility in this assay, with the remainder 'unexplained' (due to other factors).

% overlap expresses the degree to which the distributions of two groups overlap. Equal groups overlap 100%; when a treatment makes even the lowest score of the treated group higher than the highest score in the control group, overlap is 0%. In the Dulawa et al. (2004) experiment, the overlap is 69% [42, 100].

Cohen's U3 expresses the proportion of the treated group that would score above (or below) the mean for the control group. In the Dulawa et al. (2004) experiment, the samples show U3 = 0.79 [0.50, 0.95]. This means that 79% of the treated group scored below the mean for the control group, but in the population the true effect could be 50% (no effect) up to 95% (very large effect).

Probability of superiority gives the probability that any 1 random sample from group 1 will have a higher score than any 1 random sample from group 2. Values range between 0.50 (groups are equal) to 1 (no overlap between groups). In the Dulawa et al. (2004) experiment $p_{superiority}$ = 0.71 [0.50, 0.87].

N_{p80} expresses effect size in a resource-allocation framework: it is the sample-size per group needed to obtain 80% power for a subsequent study using a two-group design and an alpha of 0.05. For the Dulawa et al. (2004) experiment, N_{p80} = 26 [7, ~15.6million]. That means that if the sample is a perfect representation of the effect size, it would take 26 mice/group to have an 80% chance of again detecting the effect. Note that it is very uncertain what sample size to select for the next study—if the sample very much under-estimated the true effect it may be feasible to run the study with as few as 7 animals per group, but if the sample greatly over-estimated the true effect, it may take

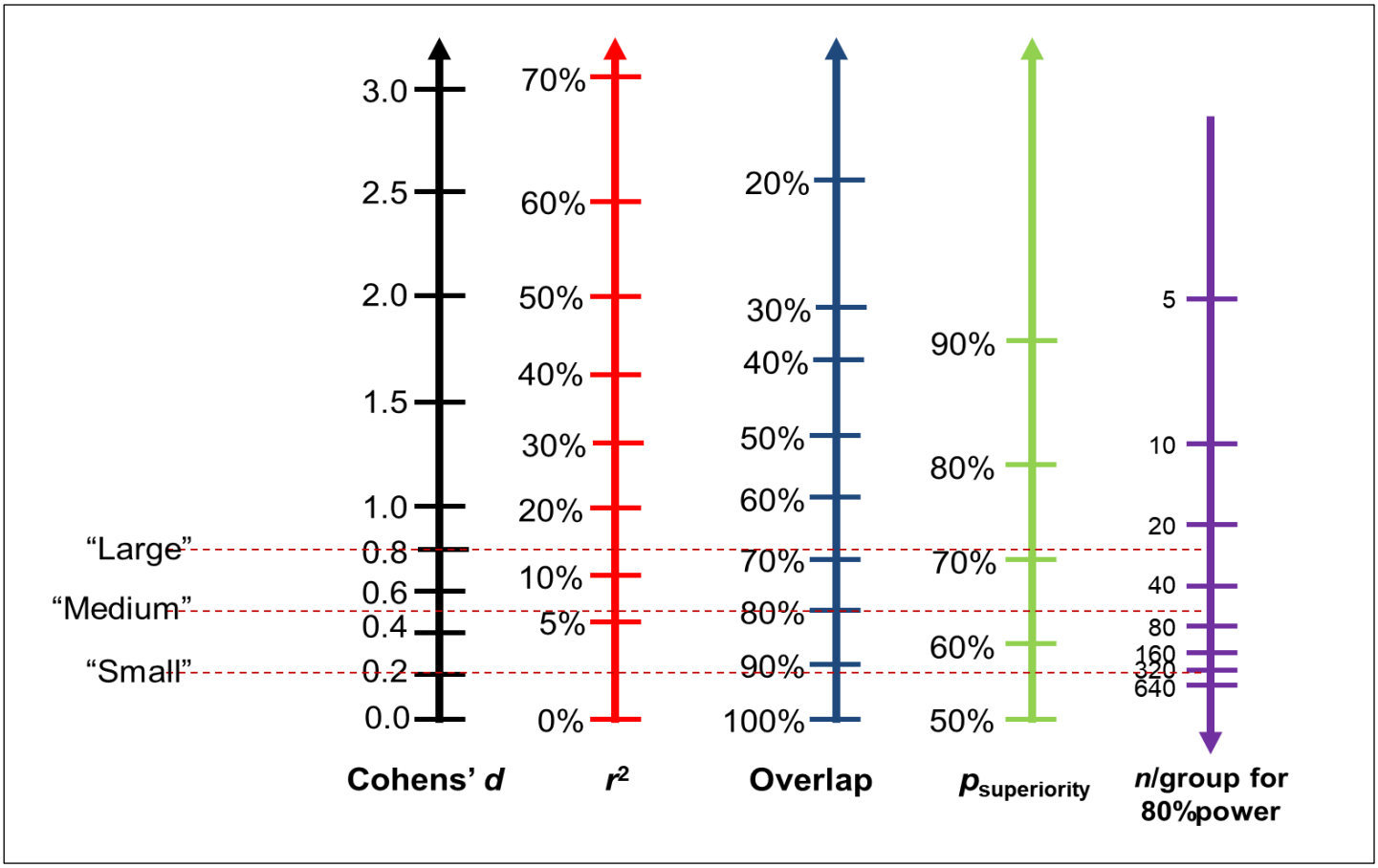

Figure 1. Comparison of some standardized Effect Size Measures. The dashed lines represent what Cohen termed large, medium, and small effects—but note that Cohen cautioned these are general rules of thumb and that researchers should judge for themselves what counts as small, medium, and large in their research domain.

millions of animals to again detect the effect. Reporting and thinking about N_{p80} is a good way to think about how replicable a result might be. When the N_{p80} is very broad, it means that the current finding provides almost no guidance in how to replicate the effect, and so it is unlikely that future researchers will select a sample size perfectly suited to the research question.

Standardized effects sizes can be useful for comparing effect sizes across different scales of measurement (e.g., comparing results across two labs that score the swim test in different ways). They are also an effective way to communicate with those who are not experts in the measurement scale. However, standardized effect sizes depend strongly on the estimated standard deviation of the dependent variable. This is problematic because: a) the sample may not provide an accurate measure of the population standard deviation; and b) the standard deviation of the dependent variable may be heterogenous across research contexts. For example, BALB/c mice might exhibit more variability in the Porsolt swim test than DBA/2 mice. Given that Cohen's d is normalized to the observed standard deviation, this can make comparing standardized effect sizes difficult. For these reasons, statisticians generally recommend reporting and interpreting raw-score effect sizes and using standardized effect size measures primarily as a supplement to aid understanding.

Why focus on effect sizes?

Practical significance. Effect sizes focus researchers on the practical significance of their findings. In Dulawa et al. (2004), considering the effect size and its confidence interval makes clear that although statistical significance was established, the data do not resolve the practical significance of how Prozac influences immobility in the Porsolt Swim test.

Adequate samples. There has been long-standing concern that the neuroscience research literature is underpowered (e.g., Button et al., 2013). Underpowered studies are ones that use a sample too small to provide accurate and reliable effect size estimates for future research. Reporting effect sizes with confidence intervals highlights the adequacy of the sample size obtained: long confidence intervals indicate inadequate data. Thus, new reporting guidelines that require reporting effect sizes with confidence intervals may help improve attention to statistical power and increase the reliability of the research literature.

Sanity Checks. Effect sizes can also function as sanity checks. For example, Elliot et al. (2007) investigated if

associations between red and danger might enable red to impair executive function. Consistent with their hypothesis, they found that a 5s exposure to red produced a statistically significant decline in verbal IQ scores. This sounds plausible given the theory. But then consider the effect size obtained; in Experiment 4 it was d = 2.2, U3 = 99%. That means that, in the sample, a brief exposure to red decreased verbal IQ by more than 2 standard deviations, equivalent to giving a normal person (IQ = 100) a mild cognitive impairment (IQ = 67). An effect size of this magnitude is simply not plausible; it is more likely an indication that something went wrong in the experiment (e.g., demand effects, poor randomization, etc.). Lakens has a fascinating discussion of how excessive effect sizes can be diagnostic of experimental woes (Lakens, 2017).

Cumulative science. Effect sizes are readily amenable to meta-analysis. Thus, they highlight opportunities to compare, combine, and synthesize across research results and contexts, something Cumming has termed 'meta-analytic thinking' (Cumming, 2011). Meta-analytic thinking helps promote science that is cumulative, integrative, and sensitive to boundary conditions.

Strategic thinking. Effect sizes also help foster strategic thinking in science. Consider, for example, a new student who could follow up two different drugs the lab has previously found to have a statistically significant effect on memory. Which project should she choose? Effect size considerations can help: the drug with the larger and more certain effect offers better chances for a fruitful second study. In this context, calculating N_{p80} is especially helpful—students can become sensitive to the resource demands of different lines of research and select projects that are actually tractable given the time and supplies available.

Optimization. When effects sizes are routinely contemplated, sample-size planning becomes routine—it simply makes sense to plan for a sample size that will be adequate for the type of effect sizes expected. Unfortunately, sample-size planning can often be daunting, indicating a need for resources far beyond what is feasible. This, however, invites the researcher to optimize the protocol to enhance the effect size. This can be done by increasing the impact of the independent variable (higher dose, more regular administration, longer administration, etc.), and/or by decreasing the noise in the measurement (within subjects design, more regular testing conditions, more homogenous participant pool, etc.).

For example, in my own research on the impact of learning on gene expression our lab started out with effect sizes of around 1.5 standard deviations (Bonnick et al., 2012). This is considered large, but still requires about 16 animals per experiment for 80% power. We optimized our protocol by switching to a within-subjects design, increasing the strength of the memory training, and regularizing our dissection protocol (Herdegen et al., 2014). We now obtain effect sizes of around 2.6 standard deviations. With much larger effect sizes we now use fewer animals (8-12/experiment) to obtain much more consistent results (Conte et al, 2017; Perez et al., 2018).

The tinkering required to really perfect an assay is at the core of the scientific method—it is the art of making the previously invisible manifest. Going through this process provides the student with deep insight into what it is that they are actually measuring. Relying on p values does not make the quality of the protocol particularly clear, in part because they are highly erratic even under ideal conditions (Lai et al., 2012). Focusing on effect sizes provides an invitation for students to always be mindful about the need to optimize measurement.

Help your students think in effect sizes

I have so far found two useful ways to help students 'think' in effect sizes. The first is to provide them with tools to help them visualize effect sizes.

- ESCI, developed by Geoff Cumming, is a free set of Excel workbooks that helps students analyze common research designs to produce graphs that strongly focus on the observed effect size and its uncertainty. ESCI also allows students to visualize data from papers in the existing literature (the **summary two** and **summary paired** tabs, for example, allow summary data from simple two-group designs to be entered and visualized). ESCI is free; it is available here: https://thenewstatistics.com/itns/esci/.
- Another wonderful tool is this interactive visualization of standardized effect size measures by Kristoffer Mangusson: http://rpsychologist.com/d3/cohend/.

What counts as a small, medium, or large effect size depends on the research context. For example, a recent review of the cognitive neuroscience literature found a wide range of reported effect sizes, with an inter-quartile range for Cohen's d of 0.64 to 1.46 (Szucs and Ioannidis, 2016). That means some cognitive neuroscientists study effects that require over a hundred participants per study (if d = 0.64, Np80 = 64/group), whereas, others are studying effects that require less than 20 participants per study (if d = 1.2, Np80 = 8/group)—that's substantial diversity!

Given the variety in what types of effect sizes are 'normal,' it is important to help students develop their own intuitions. This can be done by asking them to develop an effect size collection in a research field of interest. For example, a student interested in gender differences in the brain could look up different effect sizes in this field. They might find, for example:

- Sex differences in height are very large, around d = 1.8, N_{p80} = 6/group (https://en.wikipedia.org/wiki/Effect_size).
- Sex differences in total brain volume (not adjusted by weight) are smaller but still quite large, d = 1.1, N_{p80} = 14/group (e.g., Tan et al., 2016)
- Sex differences in rodents in the Morris Water Maze are moderate, d = 0.5, N_{p80} = 64/group (Button et al., 2013).
- Sex differences in standardized tests of math ability (e.g., the ACT) are quite subtle, d = 0.2, N_{p80} = 393/group.
- Sex doesn't always matter. For example, once adjusted for total brain volume there doesn't seem to be any sexual dimorphism in the volume of the hippocampus (Tan et al., 2016).

A collection of effect size measures like this can help the student gain some perspective on new findings and can help them evaluate which types of research projects might be most fruitful to pursue.

Further reading:

- There are a number of excellent tutorials on effect sizes (Lakens, 2013; Pek and Flora, 2017).
- Effect sizes are not just for comparing two means; they also help illuminate complex designs (Wiens and Nilsson, 2017) and can be used with dichotomous data (Haddock et al., 1998).
- Researchers are increasingly reporting their own collections of effect sizes within specific domains to make it easier to frame and interpret new results within those domains (Gignac and Szodorai, 2016; Aguinis et al., 2015).
- Two large-scale studies of low-power in the neurosciences help explain why low power is such a problem and provide useful benchmarks for different effect sizes (Button et al., 2013; Szucs and Ioannidis, 2016).

REFERENCES

Aguinis H, Field JG, Bosco FA, Aguinis H, Field JG, Pierce CA (2015) Correlational Effect Size Benchmarks 100:431–449.

American Psychological Association (2010) Publication manual of the American Psychological Association. Washington, DC.

Appelbaum M, Cooper H, Kline RB, Mayo-Wilson E, Nezu AM, Rao SM (2018) Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. Am Psychol 73:3–25.

Bonnick K, Bayas K, Belchenko D, Cyriac A, Dove M, Lass J, McBride B, Calin-Jageman IE, Calin-Jageman RJ (2012) Transcriptional Changes following long-term sensitization training and in vivo serotonin exposure in Aplysia californica. PLoS One 7:e47378.

Button KS, Ioannidis JP, Mokrysz C, Nosek B, Flint J, Robinson ESJ, Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14:365–376.

Conte C, Herdegen S, Kamal S, Patel J, Patel U, Perez L, Rivota M, Calin-Jageman RJ, Calin-Jageman IE (2017) Transcriptional correlates of memory maintenance following long-term sensitization of Aplysia californica. Learn Mem 24:502–515.

Cumming G (2011) Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis. Routledge, New York.

Cumming G, Calin-Jageman RJ (2017) Introduction to the new statistics: estimation, open science, and beyond. Routledge, New York.

Dulawa SC, Holick KA, Gundersen B, Hen R (2004) Effects of chronic fluoxetine in animal models of anxiety and depression. Neuropsychopharmacology 29:1321–1330.

Elliot AJ, Maier MA, Moller AC, Friedman R, Meinhardt J (2007) Color and psychological functioning: The effect of red on performance attainment. J Exp Psychol Gen 136:154–168.

Fidler F (2010) The American Psychological Association publication manual sixth edition: implications for statistics education. Proc ICOTS-8, Eighth Int Conf Teach Stat 8.

Gignac GE, Szodorai ET (2016) Effect size guidelines for individual differences researchers. Pers Individ Dif 102:74–78.

Haddock CK, Rindskopf D, Shadish WR (1998) Using odds ratios as effect sizes for meta-analysis of dichotomous data: a primer on methods and issues. Psychol Methods 3:339–353.

Herdegen S, Conte C, Kamal S, Calin-Jageman RJ, Calin-Jageman IE (2014) Immediate and persistent transcriptional correlates of long-term sensitization training at different CNS loci in Aplysia californica. PLoS One 9:e114481.

Lai J, Fidler F, Cumming G (2012) Subjective p intervals researchers underestimate the variability of p values over replication. Methodology 8:51–62.

Lakens D (2013) Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 4:863. doi: 10.3389/fpsyg.2013.00863.

Lakens D (2017) Impossibly hungry judges. 20% Stat 7/4/2017.

Pek J, Flora DB (2017) Reporting effect sizes in original psychological research: a discussion and tutorial. Psychol Methods doi: 10.1037/met0000126.

Perez L, Patel U, Rivota M, Calin-Jageman IE, Calin-Jageman RJ (2018) Savings memory is accompanied by transcriptional changes that persist beyond the decay of recall. Learn Mem 25:45-48. doi: 10.1101/lm.046250.117.

Picciotto M (2017) Reporting on experimental design and statistical analysis. J Neurosci 37:3737–3737.

Shirani-Mehr H, Rothschild D, Goel S, Gelman A (2016) Disentangling bias and variance in election polls. Columbia Work Pap 1–21.

Szucs D, Ioannidis JP (2016) Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. Plos Biol 15(3): e2000797. Doi: 10.1371/journal.pbio.2000797.

Tan A, Ma W, Vira A, Marwha D, Eliot L (2016) The human hippocampus is not sexually-dimorphic: meta-analysis of structural MRI volumes. Neuroimage 124:350–366.

Wiens S, Nilsson ME (2017) Performing contrast analysis in factorial designs: from NHST to confidence intervals and beyond. Educ Psychol Meas 77:690–715.

Potential Conflict of Interest Statement: Dr. Robert J Calin-Jageman is a co-author of a textbook that teaches statistics with an emphasis on effect sizes and confidence intervals.

Received March 09, 2018; revised April 04, 2018; accepted April 10, 2018.

Address correspondence to: Dr. Robert J Calin-Jageman, Psychology Department, 7900 West Division, River Forest, IL 60305. Email: rcalinjageman@dom.edu.