

TECHNICAL SPOTLIGHT

Computation of measures of effect size for neuroscience data sets

Harald Hentschke¹ and Maik C. Stüttgen²

¹Section of Experimental Anaesthesiology, Department of Anaesthesiology, University Hospital of Tübingen, Schaffhausenstraße 113, 72072 Tübingen, Germany

²Department of Biopsychology, Faculty of Psychology, University of Bochum, Germany

Keywords: confidence interval, matlab, null-hypothesis significance testing, *P*-value, statistics

Abstract

The overwhelming majority of research in the neurosciences employs *P*-values stemming from tests of statistical significance to decide on the presence or absence of an effect of some treatment variable. Although a continuous variable, the *P*-value is commonly used to reach a dichotomous decision about the presence of an effect around an arbitrary criterion of 0.05. This analysis strategy is widely used, but has been heavily criticized in the past decades. To counter frequent misinterpretations of *P*-values, it has been advocated to complement or replace *P*-values with measures of effect size (MES). Many psychological, biological and medical journals now recommend reporting appropriate MES. One hindrance to the more frequent use of MES may be their scarcity in standard statistical software packages. Also, the arguably most widespread data analysis software in neuroscience, MATLAB, does not provide MES beyond correlation and receiver-operating characteristic analysis. Here we review the most common criticisms of significance testing and provide several examples from neuroscience where use of MES conveys insights not amenable through the use of *P*-values alone. We introduce an open-access MATLAB toolbox providing a wide range of MES to complement the frequently used types of hypothesis tests, such as *t*-tests and analysis of variance. The accompanying documentation provides calculation formulae, intuitive explanations and example calculations for each measure. The toolbox described is usable without sophisticated statistical knowledge and should be useful to neuroscientists wishing to enhance their repertoire of statistical reporting.

Introduction

Since its inception in the 1920s by Ronald Fisher, the use of null-hypothesis significance testing (NHST) has pervaded much of the biological and psychological literature (Gigerenzer *et al.*, 1997). In the psychological literature, it is estimated that more than 90% of articles employ an NHST procedure (Loftus, 1991). We know of no estimate of the prevalence of NHST in neuroscience journals, but we suspect it lies in the same range. Statistically significant results are considered the prime criterion for demonstrating a treatment effect of any kind, and it is difficult, if not impossible, to publish data that fail to pass an arbitrary threshold of statistical significance (usually, $P < 0.05$ or $P < 0.01$).

Although the frequent usage of NHST is adopted by most scientists on a daily basis, there have been many articles highlighting the shortcomings of NHST in a variety of research areas, among them medicine (Fleiss, 1986; Goodman, 1999a,b), biology (Nakagawa, 2004), ergonomics (Vicente & Torenvliet, 2000), consumer research (Iacobucci, 2005) and education (Kirk, 1996, 2001; Morgan, 2003).

The misuse of NHST has perhaps been most hotly debated in psychology (Loftus, 1996; Cumming & Finch, 2005). In a special issue of *Psychological Science*, statistician John Hunter (1997) called for a ban on significance testing (also see Shrout, 1997). Following a highly influential article in *American Psychologist* (Cohen, 1994), the American Psychological Association set up a task force on statistical inference that issued its recommendations for good statistical practice (Wilkinson, 1999) and explicitly advocated standard usage of measures of effect size (MES, for a definition of the term ‘effect size’ see the paragraph further below), along with their appropriate confidence intervals (CIs). So far, their recommendations seem not to have had the desired impact as *P*-values continue to dominate statistical analysis in psychology (Fidler, 2004; Cumming *et al.*, 2007).

Among the reasons why NHST continues to be widely used (reviewed in Schmidt, 1996) may be that many programs used for statistical analysis, such as Excel and SPSS, do not provide functions to calculate MES, although a large part of them are not particularly complicated to compute. Software that does provide such functionality, e.g. scripts or packages designed for the freely available statistics program ‘R’ (Kelley, 2007; Nakagawa & Cuthill, 2007) or standalone programs (see pointers in Kline, 2004; Jordan *et al.*, 2010), seems not to be widely known or used in the neurosciences. Possibly the largest barrier to the more

Correspondence: Dr H. Hentschke, as above.
E-mail: harald.hentschke@uni-tuebingen.de

Received 31 July 2011, revised 14 September 2011, accepted 14 September 2011

widespread use of effect size statistics, in terms of software availability, is the fact that the primary language for data analysis in the neurosciences, MATLAB (The Mathworks, Natick, MA, USA), does not provide MES even in its Statistics Toolbox, with the notable exception of correlation (*corrcoef.m*, *corr.m*) and the receiver-operating characteristic (ROC) curve (*perfcurve.m*). In addition, descriptions of MES are largely scattered across the literature from diverse fields.

The aim of this article is therefore twofold: first, we want to briefly review the main criticisms of NHST and introduce measures of effect size to a neuroscience audience. The advantages of MES over the use of *P*-values are illustrated with real examples from our own research. Secondly, we describe the implementation and use of a newly developed MATLAB toolbox which provides a comprehensive set of functions for the easy calculation of parametric and non-parametric MES, together with optional bootstrapped or analytical CIs.

Shortcomings of NHST

In this section, we briefly review some of the shortcomings of NHST. For more extensive treatments, see Cohen (1994) and Loftus (1996).

In many research contexts, experimenters are confronted with the question of whether differences in some variable observed in two or more groups are 'real' (i.e. the samples were drawn from different populations) or just due to sampling error (i.e. the samples were drawn from the same distribution, and the difference in their means and standard deviations is solely due to sampling variability). The classic way to deal with this situation is to propose a null hypothesis stating that the samples were drawn from a single distribution, which is the same as saying that the difference of the means of the two populations from which samples were drawn equals zero. Assuming this 'null hypothesis' (H_0), we can calculate the probability of obtaining a difference (X) between the means of two samples from this population which is as large as or larger than that obtained (D), $P(X \geq D|H_0)$. This probability is commonly referred to as the '*P*-value'. If this probability is smaller than some predetermined value (usually, 0.05 or 0.01), the null hypothesis is considered to be unlikely and is thus rejected. In common language, the difference is termed 'statistically significant'. This logic of NHST is common to virtually all tests of significance conducted in the life sciences, regardless of whether a *t*-test, analysis of variance (ANOVA), chi-square or some non-parametric test such as Wilcoxon's signed-rank test or Mann-Whitney *U*-test are employed.

An impressive amount of empirical research has shown that many researchers employing NHST misinterpret the meaning of *P*-values. For example, Oakes (1986) found that 42 out of 70 academic psychologists questioned at a mathematical psychology meeting (incorrectly) believed that a *P*-value of 0.01 means that, if the experiment were repeated many times, 99% of repetitions would yield a significant result. Cohen (1994) termed this the 'replication fantasy'.

Tversky & Kahneman (1971) have extensively documented other misconceptions about NHST and demonstrated how they can adversely affect research. One of the most pervasive misconceptions about NHST is that the *P*-value mirrors predominantly the magnitude of an effect. However, a *P*-value is the result of several important variables, among them sample size, sample type (independent vs. dependent), type of test used and effect size. Moreover, the exact relation of these (and other) variables is rather complicated, as calculations of statistical power for an experimental design reveal (Faul *et al.*, 2007). The misconception that the smaller the *P*-value the larger the effect can be harmful to the interpretation of research. In a similar vein, a sole focus on *P*-values cannot contribute to the identification of quantitative relationships between variables, as statistically significant differences convey

information only about the direction, not the magnitude, of an effect. Building quantitative models of how experimental variables relate to each other demands other analysis methods, among them MES. As Tukey (1962) humorously put it – 'The physical scientists have learned much by storing up amounts, not just directions. If, for example, elasticity had been confined to "when you pull on it, it gets longer!", Hooke's law, the elastic limit, plasticity, and many other important topics could not have appeared...' We have summarized the four most frequent misconceptions about NHST in Table 1.

Beyond the documented existence of these misconceptions, the application of NHST suffers from frequent analysis mistakes (García-Berthou & Alcaraz, 2004; Lazic, 2010; Nieuwenhuis *et al.*, 2011), and it has been suggested that replacing NHST with an analysis couched in the framework of MES and CIs is more intuitive and hence less vulnerable to both misunderstandings and erroneous applications (Coulson *et al.*, 2010).

Definition of effect size

In this paper, we propose the use of effect size in statistical reporting. What exactly is effect size? Consider again the question of whether differences in a variable observed in two or more groups are 'real' in the sense that the underlying populations are not identical. Effect size is the magnitude of the difference between the populations. MES are statistics which quantify this difference. They come in two flavors: unstandardized and standardized. For example, mean differences belong to the category of unstandardized MES. Their magnitude is tied to the unit of measurement (a neuron's firing frequency in Hz, number of correct responses, etc.); thus, they inform us on the magnitude of effects in a way which may be preferable if we can make intuitive sense of the units of measurement. Standardized MES, in contrast, are 'metric-free' – for example, the 'd' family of MES consists of mean differences expressed in units of standard deviation of the samples. Other standardized MES metrics include correlation, proportions and ratios. It is these measures on which we focus in this article because they permit the proverbial comparison between apples and oranges by stripping samples of their units of measurements, or of differences in magnitude related to, for example, methodology. For example, field potential amplitudes in brain slice preparations are usually much higher in interface-style recording chambers than in submersion-style chambers; thus, if one were to compare the outcomes of identical experiments performed with both kinds of chambers, standardized MES would be the proper choice.

As a prime example of a measure of effect size let us consider Hedges' *g* (Hedges, 1981), applicable to the commonplace situation of two groups of data which are approximately normally distributed. Hedges' *g* (abbreviated *g*) is the difference between the means of the two groups, divided by the pooled standard deviation (Fig. 1):

$$\text{Hedges' } g = \frac{m_2 - m_1}{s_p} \quad (1)$$

where m_2 is the mean of the variable in the second (e.g. treatment) group, m_1 is the mean of the variable in the first (control) group, and s_p is the pooled standard deviation of the two groups. Thus, *g* expresses the difference between two means in a universal currency, the number of standard deviations that separate the two means (a numerical example is provided in the next paragraph). Measures of effect size exist for a range of commonly encountered analysis situations, including *n*-way analyses (up to two-way analyses in the toolbox), parametric measures for non-normal data, and data tables (Table 2, see also Fig. 5).

TABLE 1. The most common misinterpretations of null-hypothesis significance testing

Misinterpretation	Explanation	Comments
Classification fallacy	Belief that the P -value separates 'signal' ($P < 0.05$) from 'noise' ($P > 0.05$); therefore, if $P > 0.05$, there is no 'real' effect, regardless of sample size or effect size	The inductive process of NHST is asymmetric, i.e. evidence can only be gathered <i>against</i> the null hypothesis, but not in support of it; therefore, even P -values of 0.5 or higher are not indicative of the absence of an effect; put differently, absence of evidence is not evidence of absence (see also Rosenthal & Gaito, 1963, for the related 'cliff effect')
Replication fantasy	$1 - P$ is the probability to replicate the results of a study; e.g. if $P = 0.01$, the probability to replicate (obtain a significant effect under the same experimental conditions) is 0.99	The probability of replication cannot be determined on the basis of the P -value because it crucially depends on effect size and sample size as well; the probability of replication is closely connected to statistical power (Greenwald <i>et al.</i> , 1996)
Magnitude fallacy	The smaller the P -value, the larger the effect	The size of the P -value is jointly determined by the type of test (parametric vs. non-parametric), the type of the data (dependent vs. independent samples) and the magnitude of the effect; even if all factors but effect size are held constant, the relationship between effect size and P -value is highly non-linear
Illusion of attaining improbability/ wishful thinking error	The P -value denotes the probability that the null hypothesis is correct, $P(H_0)$; alternatively, the P -value denotes the probability that the null hypothesis is correct, given the data at hand, $P(H_0 D)$	As the P -value, $P(X > D H_0)$, can only be calculated under the assumption that H_0 is correct, it does not tell us anything about $P(H_0)$; the probability of interest to the researcher is, in fact, $P(H_0 D)$, which can be obtained, with several restrictions, via Bayes' theorem (Cohen, 1994; Goodman, 1999a,b; Krueger, 2001); however, the probability $P(H_0 D)$ is very much unlike the P -value, $P(X > D H_0)$; for illustration: the probability to be dead after hanging, $P(D H)$, is very much unlike the probability to have been hanged, given one is dead, $P(H D)$ (Carver, 1978)

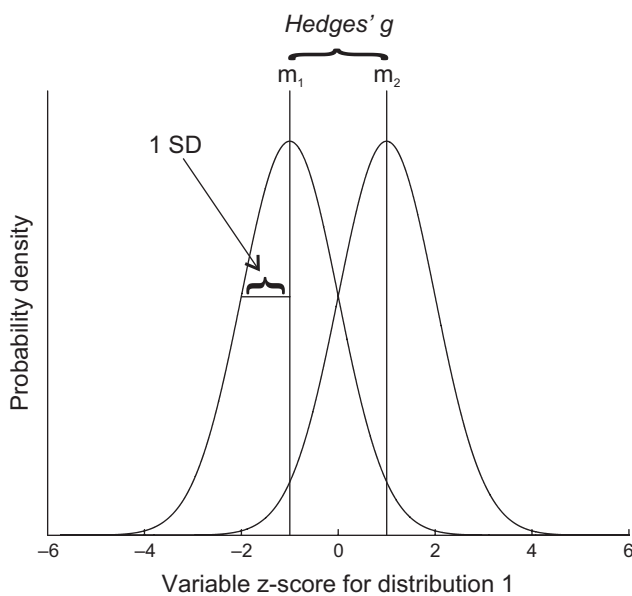


FIG. 1. Illustration of Hedges' g . Shown are two normal distributions, with means m_1 of -1 and m_2 of $+1$, and a standard deviation (SD) of 1 (identical for both distributions). Accordingly, in this example, $g = (1 - (-1))/1 = 2$.

Use and advantages of MES

In this section, we will illustrate the use and usefulness of MES for the analysis of neurophysiological data using three examples from our own work.

Example 1 – dependence of P -values but not MES on sample size

As stated in the Introduction, P -values *per se* do not convey useful information on the magnitude of an experimental effect or a group

difference. One of the underlying reasons is that P depends on other factors besides effect size, such as the number of samples. MES, by contrast, do not depend on sample size – the expected value is not a function of sample size, although the precision of the estimate is. Consider a data set obtained from mouse hippocampus. Theta oscillations were measured extracellularly *in vivo* under two conditions, before and after systemic injection of the muscarinic receptor antagonist atropine (Hentschke *et al.*, 2007). Data segments were collected from periods during which the animal did not move, corresponding to approximately 10 min recording time in each condition. As power spectra of hippocampal field potentials from non-moving animals often lack a clear theta peak, one may decide to determine the prominence of theta oscillations in the time domain, simply by measuring the amplitudes of negative-going peaks of the signal (shown for one electrode in Fig. 2A). Plugging the values obtained from one animal in a t -test for unpaired samples results in $P = 10^{-5}$, a highly significant result. However, a look at the means is sobering (Fig. 2B) – there is barely a difference between the groups, 0.018 mV in absolute terms, which is dwarfed by the standard deviation of the variable, 0.21 and 0.23 mV in the control and atropine condition, respectively. The solution to this seeming paradox is simple – both groups contain an extraordinarily large number of samples (control, $n = 6777$; atropine, $n = 5272$). This example may at first appear far-fetched – to begin with, one would usually compare the averaged peak amplitudes across animals, not thousands of individual amplitude values within animals. Furthermore, one informed look at a graph depicting the means (and, ideally, standard deviations) would dispel the illusion of a neurobiologically interesting effect of the drug treatment. Nonetheless, the example highlights a number of important points we wish to make. First, as stated before, any arbitrarily small difference between groups will eventually result in a 'significant' P -value if only sample sizes are sufficiently high, and large effects may go unnoticed in the inverse case. Although most scientists are aware of this relation when explicitly asked, there are situations in which the information necessary to judge on the relevance of P -values will not or cannot readily be gathered. MES such as Hedges' g (0.081

TABLE 2. Overview of the functions computing measures of effect size

Function	MES included in function	Argument	Complementary hypothesis test	Available CIs
<i>mes.m</i>	g_1	'g1'	One-sample <i>t</i> -test	Bootstrap
	U_{31}	'U3_1'	One-sample <i>t</i> -test	Bootstrap
	Hedges' g	'hedgesg'	Two-sample <i>t</i> -test	Bootstrap and analytical
	Glass' delta	'glassdelta'	Two-sample <i>t</i> -test	Bootstrap and analytical
	M_D/s_D	'mdbysd'	Two-sample <i>t</i> -test	Bootstrap and analytical
	$r_{\text{equivalent}}$ (point-biserial correlation)	'requiv'	Two-sample <i>t</i> -test	Bootstrap and analytical
	Common language effect size	'cles'	Two-sample <i>t</i> -test	Bootstrap
	Cohen's U_1	'U1'	Two-sample <i>t</i> -test; Mann–Whitney <i>U</i> -test	Bootstrap
	Cohen's U_3	'U3'	Two-sample <i>t</i> -test; Mann–Whitney <i>U</i> -test	Bootstrap
	AUROC	'auroc'	Two-sample <i>t</i> -test; Mann–Whitney <i>U</i> -test	Bootstrap, bootstrap <i>t</i> and analytical
<i>mes1way.m</i>	RTR, LTR	'tailratio'	Two-sample <i>t</i> -test; Mann–Whitney <i>U</i> -test	Bootstrap
	Rank-biserial correlation	'rbcorr'	Two-sample <i>t</i> -test; Mann–Whitney <i>U</i> -test	Bootstrap
	g_psi	'g_psi'	Two-sample <i>t</i> -test; any <i>post-hoc</i> test	Bootstrap and analytical
<i>mes1way.m</i>	Psibysd	'psibysd'	Two-sample <i>t</i> -test; any <i>post-hoc</i> test	Bootstrap and analytical
	Eta squared	'eta2'	One-way ANOVA	Bootstrap and analytical
	Partial eta squared	'partialeta2'	One-way ANOVA	Bootstrap and analytical
	Omega squared	'omega2'	One-way ANOVA	Bootstrap and analytical
	Partial omega squared	'partialomega2'	One-way ANOVA	Bootstrap and analytical
<i>mes2way.m</i>	g_psi	'g_psi'	Two-way ANOVA; any <i>post-hoc</i> test	Bootstrap and analytical
	Eta squared	'eta2'	Two-way ANOVA	Bootstrap
	Partial eta squared	'partialeta2'	Two-way ANOVA	Bootstrap and analytical
	Omega squared	'omega2'	Two-way ANOVA	Bootstrap
	Partial omega squared	'partialomega2'	Two-way ANOVA	Bootstrap and analytical
<i>mestab.m</i>	Risk difference	All MES computed by default	Chi-square	Analytical
	Risk ratio		Chi-square	Analytical
	Odds ratio		Chi-square	Analytical
	Phi		Chi-square	Analytical
	Sensitivity		Chi-square	None
	Specificity		Chi-square	None
	Positive predictive value		Chi-square	None
	Negative predictive value		Chi-square	None
	Binomial effect size display		Chi-square	None
	Cramer's V		Chi-square	Analytical

in the example) inform us in a straightforward way that, at less than a tenth of the combined samples' standard deviation, the effect is irrelevant. Second, the 95% CIs of (0.045 0.117) allow us to reject not only the null hypothesis of zero effect, but also all other values outside the interval (Steiger & Fouladi, 1997). Rejection of the null hypothesis in the disguise of CIs eschewing zero is probably a welcome link to familiar terrain for researchers who are unfamiliar with effect size statistics. The true merit of CIs, however, is that they facilitate a comparison of effects across studies, by defining a range of expected values of effects independent of any null hypothesis (Thompson, 2002; Nakagawa & Hauber, 2011). Third, expected values of MES such as g do not decrease with sample size (although the margin of error, expressed as CIs, does). Fourth, Hedges' g and t values relate to each other; in fact, one can be computed from the other via (here shown for independent samples):

$$g = t \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (2)$$

where t is the t -statistic from an independent samples t -test, and n_1 and n_2 are the sample sizes of the two groups.

To enlarge on these points, we randomly sampled subsets of the data, starting with an n of 10 and increasing sample size in coarsely exponential fashion up to 5000. For each pair of samples, P -values from t -tests for unpaired samples were computed, as well as Hedges' g

with the appropriate 95% confidence limits (Fig. 2C). The upper panel of Fig. 2C depicts the results of the t -tests. It can clearly be seen that P -values decrease with sample size. Statistical 'significance' ($P < 0.05$) is quite reliably obtained for sample sizes $n > 2000$, while most samples $n < 2000$ yield non-significant results, i.e. false negatives or type-II errors, due to the small effect size. By contrast, g does not depend in a systematic fashion on sample size – the values converge, with shrinking CIs, towards 0.08, the value for the full data set. The relation between P computed from t -values and g (Eqn 2) registers – all instances of $P < 0.05$ are paralleled by 95% confidence boundaries of g which exclude zero (Fig. 2C, lower panel, bold gray line and shading).

The finding of such a small effect size carries important implications for researchers aiming to replicate or extend results in a similar situation. An *a priori* power analysis (conducted with G*Power 3.1, Faul *et al.*, 2007) reveals that, to replicate the finding with a power of 0.8 (i.e. to obtain a significant difference between atropine and control conditions with a probability of 80% using a t -test) requires a minimum total sample size of 4788 data points. If one wishes to be on the safe side and have a power of 0.95, the sample size should be increased to 7926 data points. A researcher who is ignorant of these facts may therefore repeat the experiment but record for only, say, 2 min and obtain roughly 1000 data points. The probability of replicating this effect would be merely 25% for a two-tailed t -test and 36% for a one-tailed t -test (if in the right direction). Thus, the

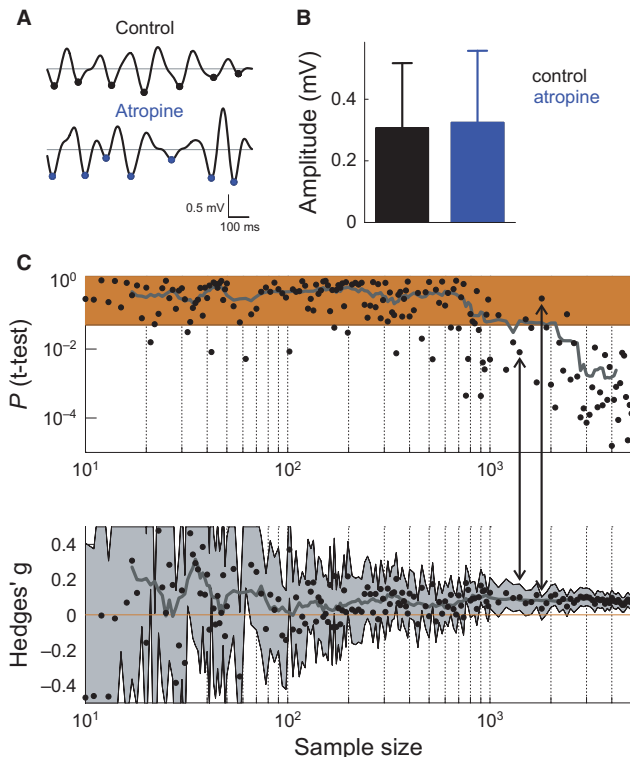


FIG. 2. Influence of sample size on P -values and Hedges' g . (A) Short excerpts of field potential recordings in mouse hippocampus *in vivo*, filtered to isolate theta oscillations (5–12 Hz). The amplitudes of theta troughs under control conditions and after injection of atropine, marked by black and blue dots, respectively, were tested against each other. (B) Means and standard deviation of both samples, collected over approximately 10 min of recording time. (C) Values of P (top panel) and Hedges' g (bottom panel) for randomly sampled subsets of the full data set. In the top panel, the range of P -values classically associated with the absence of an effect is marked in red (note logarithmic scaling of abscissa and ordinate). In the bottom panel, the gray area represents the 95% confidence interval of Hedges' g (clipped at low sample sizes) and the 'no effect' value of zero is marked by the red dotted line. In both graphs, the gray line is a running average of eight adjacent values, illustrating that P -values, but not Hedges' g , decrease with sample size. Note that $P < 0.05$ corresponds to confidence intervals excluding zero (compare the two specific random samples marked by double arrows). For interpretation of color references in figure legend, please refer to the Web version of this article.

researcher is likely to fail to replicate this finding and may discuss this 'failure' away with different mouse strains or different electrode positions, even though it is merely a result of sampling variability and the binary nature of decisions dictated by common practice of NHST. By contrast, as inspection of the bottom panel of Fig. 2C reveals, even with only 1000 data points, that Hedges' g would likely fall in a similar range as in the previous scenario, 0–0.2. Thus, what seems at first glance to be a failure to replicate turns out to be quite replicable in terms of measurement of effect size.

Example 2 – identifying discrete cell populations on the basis of spike responses

The rat trigeminal ganglion (TG) contains a population of mechanosensitive neurons which respond to the mechanical stimulation of a single whisker on the animals' muzzle (Zucker & Welker, 1969; Stüttgen *et al.*, 2006). Several studies have shown that these neurons' firing rates covary with kinematic parameters of the whisker deflection, such as amplitude and velocity (Gibson & Welker, 1983). We were

interested to see whether these neurons form discrete, i.e. non-overlapping subgroups, each coding for only one kinematic parameter. We recorded the activity of single neurons from the TG in anesthetized rats while applying a parameterized stimulus set which included 15 stimuli, made up of all possible combinations of three different deflection amplitudes and five different deflection velocities (Stüttgen *et al.*, 2008). Each stimulus was presented 10 times in pseudorandom order. A standard analysis for this kind of data is to count spike responses to each stimulus and to subject the data to a two-way 3*5 factorial ANOVA with amplitude and velocity as factors. We did so for each of 22 neurons individually, and it turned out that 21/22 neurons had statistically significant response modulations due to amplitude ($P < 0.05$), 20/22 had significant modulations due to velocity and 8/22 showed a significant interaction of these factors. These numbers were essentially unchanged when the significance criterion was set to 0.01 (the respective numbers are 21, 19 and 7). On the basis of this result, one may conclude that most TG neurons code for both velocity and amplitude, and that some show more complex response properties, being modulated by a multiplicative interaction of the two factors.

Although this statement is formally correct, there has been no mention of the *degree* to which these neurons are modulated by the kinematic parameters. As outlined in the Introduction, the P -value does not allow an inference as to the relative magnitude of the effects. One measure of effect size appropriate for this kind of analysis design is called η^2 :

$$\eta^2 = \frac{SS_{\text{effect}}}{SS_{\text{total}}} \quad (3)$$

where SS_{effect} is the sum of squares between groups (treatments) and SS_{total} is the overall sum of squares. The intuition behind η^2 is: how much of the total variance in the data is explained by any one factor, i.e. amplitude, velocity or their interaction? In the present design, three kinds of η^2 can be computed, one for each of the above factors. The sum of all three η^2 cannot exceed 1 – if the sum equaled 1, all the variance in the data would be explained by the experimental factors, which is rarely encountered. Values of ≈ 0.1 , ≈ 0.2 and ≥ 0.3 may be roughly classified as small, medium and large effects, corresponding to 10, 20 and 30% of explained variance, but note that verbal judgments such as 'large' may vary substantially across and within research fields.

We calculated η^2 for amplitude and velocity for each neuron (η^2 for the interaction was negligible). The resulting scatterplot is depicted in Fig. 3. It is obvious that there are two separable clusters of neurons,

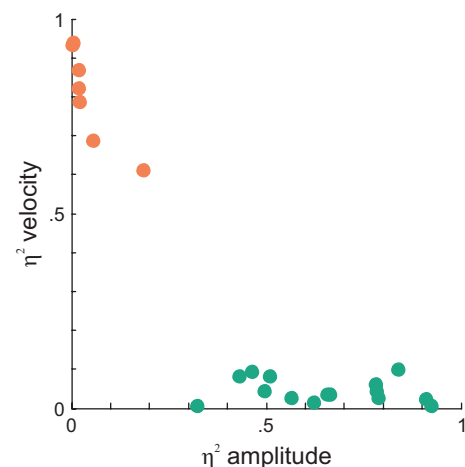


FIG. 3. Measures of effect size (η^2) for amplitude and velocity for 22 TG neurons (figure modified after Stüttgen *et al.*, 2008).

those with large (> 0.3) values for amplitude (green) and with large values for velocity (orange). Recall that almost all of these neurons were significantly modulated by both amplitude and velocity. However, plotting η^2 for these factors against each other reveals that every neuron is modulated primarily by one of these factors and to a much lesser extent by the other. For this data set, 42/66 P -values were smaller than 0.001 and 38/66 were even smaller than 0.000001. Thus, the size of the P -value carries hardly any implication about the magnitude of an effect, and the exclusive use of NHST for these data would have concealed a straightforward relationship between kinematic parameters and neural response types. Instead, one would have opted erroneously for a considerably more complex scheme to explain neural coding of kinematic events at the level of primary afferents.

Example 3 – comparing peri-event time histograms (PETHs)

Multiunit action potentials were recorded from organotypic cultures of mouse neocortex in two different formulations of artificial cerebrospinal fluid (ACSF1 and ACSF2), which differed in the concentrations of magnesium, calcium and potassium. The question was whether action potential firing over the course of spontaneous UP states differed between the two conditions. To this end, the onsets of UP states were determined on the basis of the field potential (recorded with the same electrode), and action potentials were collected in PETHs around those time points, converted to firing frequency and averaged for each recording (ACSF1, $n = 64$ recordings; ACSF2, $n = 39$). Figure 4A depicts the median PETHs for both conditions. The median firing rates obviously do not change randomly from bin to bin, but rather follow a smooth course. A good approach would be to fit a

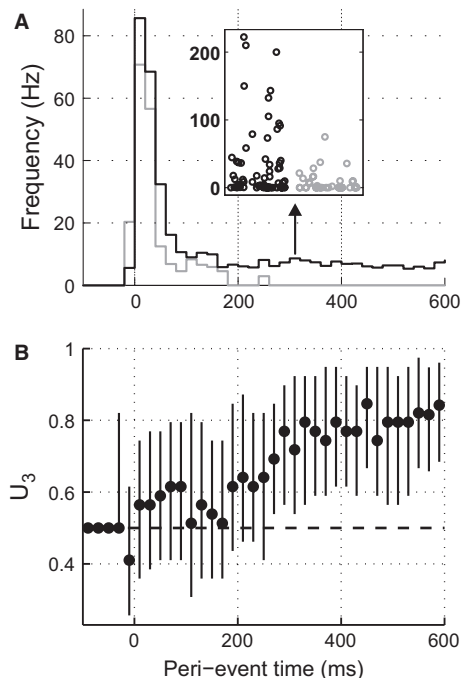


FIG. 4. Comparing peri-event time histograms (PETHs) of spikes with Cohen's U_3 , a non-parametric measure of effect size. (A) Graphs representing median PETHs of extracellular multiunit recordings from *in vitro* preparations of mouse neocortex under control (black, $n = 64$) and treatment (gray, $n = 39$) conditions. Inset – the individual data for the bin at 300–320 ms, illustrating their non-normal distribution (same color code as for the PETHs; abscissa offset is arbitrary). (B) Bin-by-bin values of U_3 (dots) and 95% confidence intervals (lines) obtained by bootstrapping. Dotted line at ordinate value of 0.5 corresponds to zero effect.

function to the data; for example, exponentials could be fitted to the phases of decaying firing rates and the fitted parameter(s) used to judge on the difference between the conditions (Motulsky & Christopoulos, 2003). However, PETHs often take more complicated forms, making the choice of the function to fit difficult. An alternative might be a factorial analysis (ANOVA and/or matching MES) with treatment (ACSF type) and time (bins) as the factors. However, both the fitting and the factorial approaches would yield doubtful results for this data set due to the highly non-normal distribution of the firing frequencies in each time bin, which have a strong bias toward values at or around zero (see Fig. 4A, inset). This might be mitigated by a transformation of the data which bestowed approximate normality on them. A simpler approach, practically devoid of assumptions on the nature of the data, is based on a non-parametric measure of effect size, Cohen's U_3 (Cohen, 1988). U_3 is the proportion of data points in the 'lower' group (here, ACSF2, gray in Fig. 4A) which are smaller than the median of the 'higher' group (ACSF1; see the toolbox' documentation for more information on U_3). The results of the computations are presented in Fig. 4B. For the first 260 ms post-event, U_3 hovers mostly between 0.5, the zero effect value, and 0.6. Beyond 260 ms, the proportion of scores grows more or less continuously, up to over 80% in the last bin. With a similar trend, the 95% CIs of U_3 , obtained by bootstrapping, recede from the zero effect line, confirming the notion of a more sustained firing in the first condition. Thus, via a bin-wise comparison of firing rates we retain temporally resolved information on the effects of the treatment. A similar reasoning, and motivation to use MES as primary statistics or complements, applies to, for example, recordings from closely spaced electrodes (Hentschke *et al.*, 2009) and other data sets within which the variable of interest varies smoothly. P -values would reduce the issue at hand to a series of dichotomous decisions. The display of effect size including confidence intervals, by contrast, provides us with the appropriate view of graded effects.

The MES toolbox

Nakagawa & Cuthill (2007) argue that three types of effect statistics suffice for most situations: r statistics (correlation coefficients such as Pearson's product-moment coefficient, Spearman's rank correlation, point-biserial and phi), d statistics (Cohen's d or Hedges' g) and the odds ratio [one of three most used comparative risk measurements, which, according to Nakagawa & Cuthill (2007), are odds ratio, relative risk and risk difference]. These three types of MES are particularly useful because they can be computed for pair-wise comparisons, even if the data design is factorial (ANOVA-type). However, in a factorial design our primary interest often lies in overall (omnibus) effects of the underlying factors (Example 2 above) or in focused comparisons between more than two groups. The comparatively well-known two-sample variants of these measures are not appropriate or applicable in this situation. Furthermore, samples may not be normally distributed (Example 3). Accordingly, we included MES such as η^2 , ω^2 , MES for contrasts and a selection of non-parametric effect size indices such as Cohen's U_1 and U_3 and the area under the ROC curve.

The MES toolbox is written in MATLAB and requires the Statistics Toolbox. It can be downloaded from two sources, <http://sourceforge.net/projects/mestoolbox/> and the MATLAB Central File Exchange, <http://www.mathworks.com/matlabcentral/fileexchange/32398-measures-of-effect-size-toolbox>. It is accompanied by example data and an extensive documentation and reference manual. All MES have been condensed into four functions: *mes.m* for two-sample designs, *mes1way.m* for one-way ANOVA designs, *mes2way.m* for two-way ANOVA designs and *mestab.m* for data sets with absolute frequencies.

By specifying optional arguments, the user can fully exploit the entire repertoire and functionality of MES that the toolbox has to offer. This includes the option to specify whether (or across which factor) the data are dependent (repeated measures), but note that a few MES are only defined for either dependent or independent data or implemented only for one of these cases.

Perhaps most importantly, CIs can be computed for all MES, with the exception of five MES included in *mestab.m*. For many MES, different types of CIs are available. Those based on bootstrapping are implemented for all MES contained in *mes.m*, *mes1way.m* and *mes2way.m*; the number of iterations can be specified by the user. For a range of MES, analytical CIs may be computed. For a subset among these, there is a choice among 'approximate' and 'exact' intervals. Approximate intervals are based on central χ^2 -, t - and F -distributions, which also underlie the computation of P -values in NHST. Exact CIs are computed in an iterative way from the so-called non-centrality parameter (ncp) of non-central χ^2 -, t - and F -distributions (Steiger & Fouladi, 1997; Thompson, 2002; Smithson, 2003; Kelley, 2007). Formulae for a number of MES to convert ncp to CIs were taken from these studies, as well as from Fidler & Thompson (2001). We also wish to point out the possibility to compute standardized contrasts in one- and two-way designs. As has been stated before (Rosenthal *et al.*, 2000; Kline, 2004; Nakagawa & Cuthill, 2007), in many situations the overall effect of a treatment on several experimental groups is of secondary interest. Contrast analysis permits the computation of MES for focused comparisons, for example of one treatment group vs. the weighted average of two other groups.

Table 2 provides an overview over the four functions, the MES they provide, the related type of hypothesis test and the type(s) of CIs available. All functions return a structure 'stats' as output. 'stats' contains fields holding the value of the requested MES, sample size, analysis design (paired vs. unpaired), number of bootstrapping iterations, confidence level, CIs and the computational method underlying the CI. Furthermore, as a number of MES are computed from the same terms that are needed to generate t - and F -values of t -tests and ANOVAS, respectively, *mes.m* produces t -statistics and *mes1way.m* and *mes2way.m* produce full ANOVA results tables.

Figure 5 provides a structured overview over the MES featured in the toolbox, which may aid researchers in the selection of appropriate MES for their specific type of question. More extensive coverage of each MES, along with illustrations and example computations, are provided in the documentation of the MES toolbox. The documentation may be informative even for readers not familiar with MATLAB, because it provides intuitive explanations, calculation formulae and guidance for the interpretation of each MES covered.

In summary, we strongly believe that data analysis in the neurosciences would benefit from the standard computation of MES. We have provided three examples in which MES reveal important aspects of data which would have been undetected, underappreciated or dealt with in awkward ways with standard NHST. Another benefit of using MES is meta-analysis. Nakagawa & Hauber (2011) have recently argued that the neurosciences may benefit from a shift in focus, from individual studies to a meta-analytic perspective. Distilling statistical results through meta-analysis has proved immensely useful in several fields of science, such as psychology and especially medicine (e.g. Smith & Glass, 1977; Mann, 1990; Baigent *et al.*, 2005), and MES are central to that endeavor (Thompson, 2002). We therefore feel that editorial policies encouraging or demanding the routine use of MES (as already implemented in several psychological and medical journals) would be advantageous. Nonetheless, we agree with Sarter & Fritschy (2008) that partisan stances are not helpful; in particular, a ban of NHST would probably not bear fruit, but rather provoke bearish responses. However, at a minimum, authors should always specify sufficiently ample detail of their data and their statistical analyses, such that an assessment of effect size is possible, and refrain from misuses of NHST, such as claiming the absence of an effect on the basis of low sample sizes. Many authors do follow this practice, and guidelines to this effect are already in place for the *European Journal of Neuroscience* (Sarter & Fritschy, 2008). On top of this, we reiterate the appeal to our colleagues to open their minds to an alternative kind of statistics. Our hope is that the availability of software allowing the easy and flexible calculation of MES will aid in this endeavor, eventually improving the presentation and discussion of neuroscientific results.

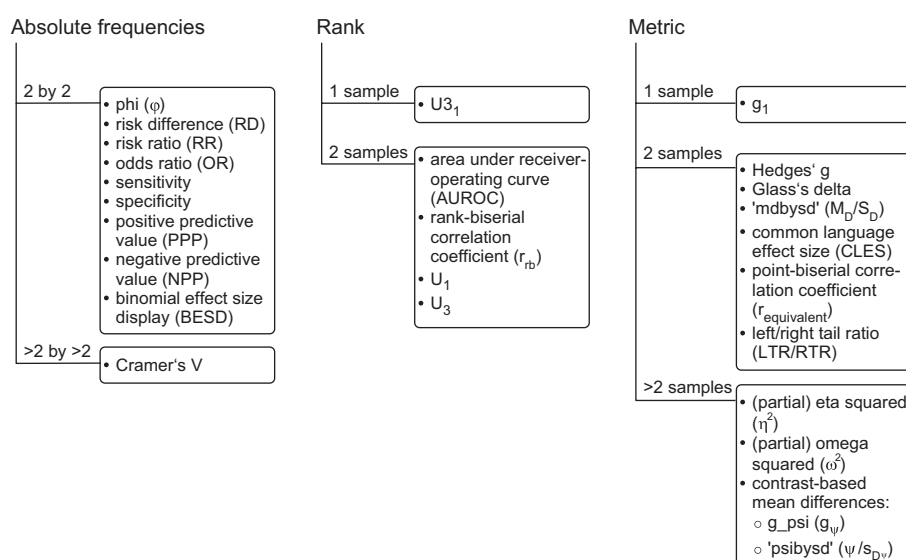


FIG. 5. Decision tree for choosing an appropriate measure of effect size available in the toolbox. Note that several aspects of the data which must guide the decision do not show up in this figure (among them the presence or absence of within-subjects factors and normality of the data); for more detailed information see Table 2 and the documentation accompanying the toolbox.

Acknowledgements

We thank Luise Liebig for providing unpublished data and Robert Pearce, Matthew I. Banks and Cornelius Schwarz for agreeing to the use of published data sets in the examples.

Abbreviations

ACSF, artificial cerebrospinal fluid; ANOVA, analysis of variance; CI, confidence interval; MES, measures of effect size; NHST, null-hypothesis significance testing; PETH, peri-event time histogram; ROC, receiver-operating characteristic; TG, trigeminal ganglion.

References

- Baigent, C., Keech, A., Kearney, P.M., Blackwell, L., Buck, G., Pollicino, C., Kirby, A., Sourjina, T., Peto, R., Collins, R. & Simes, J. (2005) Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90056 participants in 14 randomised trials of statins. *Lancet*, **366**, 1267–1278.
- Carver, R.P. (1978) The case against statistical significance testing. *Harv. Educ. Rev.*, **48**, 378–399.
- Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, New York, NY.
- Cohen, J. (1994) The earth is round ($P < .05$). *Am. Psychol.*, **49**, 997–1003.
- Coulson, M., Healey, M., Fidler, F. & Cumming, G. (2010) Confidence intervals permit, but do not guarantee, better inference than statistical significance testing. *Front. Psychol.*, **1**, 26, doi: 10.3389/fpsyg.2010.00026.
- Cumming, G. & Finch, S. (2005) Inference by eye: confidence intervals and how to read pictures of data. *Am. Psychol.*, **60**, 170–180.
- Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., Lo, J., McMenamin, N. & Wilson, S. (2007) Statistical reform in psychology: is anything changing? *Psychol. Sci.*, **18**, 230–232.
- Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007) G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods*, **39**, 175–191.
- Fidler, F. (2004) Editors can lead researchers to confidence intervals, but can't make them think. *Psychol. Sci.*, **15**, 119–125.
- Fidler, F. & Thompson, B. (2001) Computing correct confidence intervals for anova fixed- and random-effects effect sizes. *Educ. Psychol. Meas.*, **61**, 575–604.
- Fleiss, J.L. (1986) Confidence intervals vs. significance tests: quantitative interpretation. *Am. J. Public Health*, **76**, 587.
- García-Berthou, E. & Alcaraz, C. (2004) Incongruence between test statistics and *P* values in medical papers. *BMC Med. Res. Methodol.*, **4**, 13.
- Gibson, J.M. & Welker, W.I. (1983) Quantitative studies of stimulus coding in first-order vibrissa afferents of rats. 1. Receptive field properties and threshold distributions. *Somatosens. Res.*, **1**, 51–67.
- Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. & Krüger, L. (1997) *The Empire of Chance: How Probability Changed Science and Everyday Life*. Cambridge University Press, Cambridge.
- Goodman, S.N. (1999a) Toward evidence-based medical statistics. 1: the *P* value fallacy. *Ann. Intern. Med.*, **130**, 995–1004.
- Goodman, S.N. (1999b) Toward evidence-based medical statistics. 2: the Bayes factor. *Ann. Intern. Med.*, **130**, 1005–1013.
- Greenwald, A.G., Gonzalez, R., Harris, R.J. & Guthrie, D. (1996) Effect sizes and *P* values: what should be reported and what should be replicated? *Psychophysiology*, **33**, 175–183.
- Hedges, L.V. (1981) Distribution theory for Glass's estimator of effect size and related estimators. *J. Educ. Behav. Stat.*, **6**, 107–128.
- Hentschke, H., Perkins, M.G., Pearce, R.A. & Banks, M.I. (2007) Muscarinic blockade weakens interaction of gamma with theta rhythms in mouse hippocampus. *Eur. J. Neurosci.*, **26**, 1642–1656.
- Hentschke, H., Benkowitz, C., Banks, M.I., Perkins, M.G., Homanics, G.E. & Pearce, R.A. (2009) Altered GABA_A slow inhibition and network oscillations in mice lacking the GABA_A receptor β_3 subunit. *J. Neurophysiol.*, **102**, 3643–3655.
- Hunter, J.E. (1997) Needed: a ban on the significance test. *Psychol. Sci.*, **8**, 3–7.
- Iacobucci (2005) On *P*-values. *J. Consum. Res.*, **32**, 6–11.
- Jordan, D., Steiner, M., Kochs, E.F. & Schneider, G. (2010) A program for computing the prediction probability and the related receiver operating characteristic graph. *Anesth. Analg.*, **111**, 1416–1421.
- Kelley, K. (2007) Confidence intervals for standardized effect sizes: theory, application, and implementation. *J. Stat. Softw.*, **20**, 1–24.
- Kirk, R.E. (1996) Practical significance: a concept whose time has come. *Educ. Psychol. Meas.*, **56**, 746–759.
- Kirk, R.E. (2001) Promoting good statistical practices: some suggestions. *Educ. Psychol. Meas.*, **61**, 213–218.
- Kline, R.B. (2004) *Beyond Significance Testing*. American Psychological Association, Washington, DC.
- Krueger, J. (2001) Null hypothesis significance testing. On the survival of a flawed method. *Am. Psychol.*, **56**, 16–26.
- Lazic, S.E. (2010) The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neurosci.*, **11**, 5.
- Loftus, G.R. (1991) On the tyranny of hypothesis testing in the social sciences. *Contemp. Psychol.*, **36**, 102–105.
- Loftus, G.R. (1996) Psychology will be a much better science when we change the way we analyze data. *Curr. Dir. Psychol. Sci.*, **5**, 161–171.
- Mann, C. (1990) Meta-analysis in the breach. *Science*, **249**, 476–480.
- Morgan, P.M. (2003) Null hypothesis significance testing: philosophical and practical considerations of a statistical controversy. *Exceptionality*, **11**, 209–221.
- Motulsky, H.J. & Christopoulos, A. (2003) *Fitting Models to Biological Data Using Linear and Nonlinear Regression. A Practical Guide to Curve Fitting*. Oxford University Press, New York, NY.
- Nakagawa, S. (2004) A farewell to Bonferroni. *Behav. Ecol.*, **15**, 1044–1045.
- Nakagawa, S. & Cuthill, I.C. (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol. Rev.*, **82**, 591–605.
- Nakagawa, S. & Hauber, M.E. (2011) Great challenges with few subjects: statistical strategies for neuroscientists. *Neurosci. Biobehav. Rev.*, **35**, 462–473.
- Nieuwenhuis, S., Forstmann, B.U. & Wagenmakers, E.-J. (2011) Erroneous analyses of interactions in neuroscience: a problem of significance. *Nat. Neurosci.*, **14**, 1105–1107.
- Oakes, M. (1986) *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. Wiley, New York, NY.
- Rosenthal, R. & Gaito, J. (1966) The interpretation of levels of significance by psychological researchers. *J. Psychol.*, **55**, 33–38.
- Rosenthal, R., Rosnow, R.L. & Rubin, D.B. (2000) *Contrasts and Effect Sizes in Behavioral Research: A Correlational Approach*. Cambridge University Press, Cambridge.
- Sarter, M. & Fritschy, J.-M. (2008) Reporting statistical methods and statistical results in E.J.N. *Eur. J. Neurosci.*, **28**, 2363–2364.
- Schmidt, F.L. (1996) Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychol. Methods*, **1**, 115–129.
- Shrout, P.E. (1997) Should significance tests be banned? *Psychol. Sci.*, **8**, 1–2.
- Smith, M.L. & Glass, G.V. (1977) Meta-analysis of psychotherapy outcome studies. *Am. Psychol.*, **31**, 752–760.
- Smithson, M. (2003) *Confidence Intervals*. Sage Publications, Thousand Oaks, CA.
- Steiger, J.H. & Fouladi, R.T. (1997) Noncentrality interval estimation and the evaluation of statistical models. In Harlow, L.L., Mulaik, S.A. & Steiger, J.H. (Eds), *What if There were no Significance Tests?* Erlbaum, Mahwah, NJ, pp. 221–257.
- Stüttgen, M.C., Rüter, J. & Schwarz, C. (2006) Two psychophysical channels of whisker deflection in rats align with two neuronal classes of primary afferents. *J. Neurosci.*, **26**, 7933–7941.
- Stüttgen, M.C., Kullmann, S. & Schwarz, C. (2008) Responses of rat trigeminal ganglion neurons to longitudinal whisker stimulation. *J. Neurophysiol.*, **100**, 1879–1884.
- Thompson, B. (2002) What future quantitative social science research could look like: confidence intervals for effect sizes. *Educ. Researcher*, **31**, 25–32.
- Tukey, J.W. (1962) The future of data analysis. *Ann. Math. Statist.*, **33**, 1–67.
- Tversky, A. & Kahneman, D. (1971) Belief in the law of small numbers. *Psychol. Bull.*, **76**, 105–110.
- Vicente, K.J. & Torenvliet, G.L. (2000) The earth is spherical ($P < .05$): alternative methods of statistical inference. *Theor. Issues Ergon. Sci.*, **1**, 248–271.
- Wilkinson, L. (1999) Statistical methods in psychology journals. *Am. Psychol.*, **54**, 594–604.
- Zucker, E. & Welker, W.I. (1969) Coding of somatic sensory input by vibrissae neurons in the rat's trigeminal ganglion. *Brain Res.*, **12**, 138–156.