BRAIN IMPAIRMENT VOLUME 19 NUMBER 1 MARCH pp. 70–80 © Australasian Society for the Study of Brain Impairment 2017
doi:10.1017/BrImp.2017.28

Null Hypothesis Significance Testing, *p*-values, Effects Sizes and Confidence Intervals

Between-groups research

Michael Perdices

Department Of Neurology, Royal North Shore Hospital, New South Wales, Australia

There has been controversy over Null Hypothesis Significance Testing (NHST) since the first quarter of the 20th century and misconceptions about it still abound. The first section of this paper briefly discusses some of the problems and limitations of NHST. Overwhelmingly, the 'holy grail' of researchers has been to obtain significant *p*-values. In 1999 the American Psychological Association (APA) recommended that if NHST was used in data analysis, then researchers should report effect sizes (ESs) and their confident intervals (CIs) as well as *p*-values. The APA recommendations are summarised in the next section of the paper. But as neuropsychological rehabilitation clinicians, the primary interest is (or should be) to determine whether or not the effect of an intervention is clinically important, not just statistically significant. In this context, ESs and their CIs provide information relevant to clinicians. The next section of the paper reviews common ESs and worked out examples are provided for the calculation of three commonly used ES (Cohen's ***d***, Hedge's ***g*** and Glass' *delta*). Web-based resources for calculating other ESs and their CIs are also reviewed.

Keywords: NHST, Effect Size, Confidence Interval, *p*-values

Null Hypothesis Significance Testing and *p*-values

Ronald Aylmer Fisher (1925) established the foundations of Null Hypothesis Significance Testing (NHST). Subsequently, Neyman and Pearson, (1928a, 1928b) proposed a slightly different approach. The point of departure for Fisher's approach was to start from the population and determine the probability that the sample (i.e., the group being studied) had been drawn from a given population. It required only one hypothesis: the null hypothesis (H_0), usually stated as there being no statistically significant difference between two groups on the variable being investigated. Neyman and Pearson's approach (which Fisher vehemently opposed) was to start from the sample and determine the probability that it was from a given population that the sample was drawn. This approach required an alternative, or experimental hypothesis (H_1), usually stated as there being a statistically significant difference between two groups on the variable being investigated, and against which the null hypothesis would be evaluated. Fisher's approach has been regarded as significance testing and that of Neyman and Pearson as null hypothesis testing (Huberty & Pike, 1999). Over the years, reasoning in statistical inference adopted in behavioural research, and still in current use, is an amalgam of both (Falk & Greenbaum, 1995). Gigerenzer (1993) attributes the confusion and misconceptions that plague NHST to this hybridisation.

The 'hybrid logic' of NHST is usually applied as follows. We might wish to investigate the effectiveness of an intervention aimed to improve the ability of patients with stroke to remember

Address for correspondence: Department Of Neurology, Royal North Shore Hospital, The University of Sydney Medical School, Northern Clinical School, Discipline of Psychiatry, New South Wales, Australia.
E-mail: Michael.Perdices@health.nsw.gov.au

https://doi.org/10.1017/BrImp.2017.28 Published online by Cambridge University Press

appointments. In order to do so, two randomly drawn samples of patients could be studied: an experimental sample of who received the intervention and a control sample of those who did not. The efficacy of the intervention could then be examined by comparing the mean number of appointments remembered by the treated group ($\overline{X}_T$) and the control group ($\overline{X}_C$). The null hypothesis would generally be expressed as H_0: $\overline{X}_T = \overline{X}_C$, and the alternative or experimental hypothesis as H_1: $\overline{X}_T \neq \overline{X}_C$ if two-tailed or H_1: $\overline{X}_T > \overline{X}_C$ if one-tailed. By convention, the criterion for significance, α, is usually set at 0.05 (it is worth keeping in mind that Fisher chose the 0.05 level without providing any theoretical basis or rationale). An independent samples *t*-test could be used to analyse the data (note that the same reasoning applies if any other statistical test is used). If the *p*-value yielded by the *t*-test was <0.05, a common belief is that there would be less than a 5% chance of H_0 being true and would confidently reject it. We have proven that the treatment was effective. Seen from this perspective, NHST provides neat and unambiguous answers, hence its seductiveness. But closer scrutiny shows that our understanding of the process and interpretation of the results of statistical tests is, probably, rather muddled.

Berkson (1938) was one of the first to discuss the shortcomings of the NHST, and there has been lively debate ever since (e.g., Carver, 1978; Clark, 1963; Cohen, 1990; Glaser, 1999; Gliner, Leech & Morgan, 2002; Meehl, 1967; Nickerson, 2000; Rozeboom, 1960). Coming to grips with NHST requires a clear understanding of the relation between populations and samples, the difference between the sampling distribution and the population distribution, the postulates of the law of large numbers and the central limit theorem. It also requires a clear understanding of the meaning of significance level, H_0, H_1 and *p*-values, as well as the interrelationship between them (Castro Sotos, Vanhoof, Van den Noortgate & Onghena, 2007). However, misconceptions about NHST abound (Krishnan & Idris, 2014; Lambdin, 2012; Vallecillos & Batanero, 1997a, 1997b; Vallecillos, 2001).

An overarching problem with NHST is that it does not do what we want it to do (Cohen, 1990: Cohen, 1994; Falk & Greebaum, 1995; Kirk, 1996). What the researcher wants to know is the probability of H_0 being true given the observed data D; this can be expressed as $P(H_0|D)$. What NHST gives us is the probability of obtaining the observed data if H_0 is true, which can be expressed as $P(D|H_0)$. Clearly, $P(H_0|D)$ is not the same as $P(D|H_0)$. If the statistical test employed yields significant results (i.e., $p \leq 0.05$) all it tells us is that $P(D|H_0)$ is low. This does not necessarily mean that $P(H_0|D)$ is also low and consequently it is erroneous to conclude that a significant result means that H_0 is false. Carver (1978) provided a very clear illustration of this. He suggested that the probability of a person being dead (symbolised as D), if they had been hanged (symbolised as H) can be expressed as $P(D|H)$. According to Carver, this probability would be very high. The reverse, the probability that a person has been hanged, given that the person is dead can be expressed as $P(H|D)$. However, it is abundantly clear that $P(H|D)$ is likely to be very small. Carver points out that in his example it is unlikely that anyone would mistakenly substitute $P(D|H)$ for $P(H|D)$. By analogy, this is what occurs in NHST.

What Should Be Done?

In 1994, Cohen published a seminal article reviewing the problems and misunderstandings that had beleaguered NHST over the previous four decades. Following the article's publication, the Board of Scientific Affairs of the American Psychological Association (APA) appointed the Taskforce on Statistical Inference (APA-TSI) to consider the role of NHST in psychological research and make recommendations regarding analysis of quantitative research in psychology (Wilkinson et al., 1999). There had been expectations in some quarters that there would be an outright ban on NHST (Kirk, 1996). This did not occur and the Taskforce made several recommendations that researchers should implement in conjunction with NHST and reporting of *p*-values. These are summarised in Table 1.

In a nutshell, the main focus of the Taskforce recommendations was that results of NHST should report exact *p*-values and not just dichotomous decisions such as accept/reject H_0. More importantly, it was recommended that in addition to *p*-values, effect size (ES) should also be reported, as well as the precision or confidence interval (CI) of the ES. These recommendations were incorporated in subsequent editions of the APA Publication Manual. Prior to the 6th edition of the Publication Manual, the Working Group on Journal Article Reporting Standards (JARS) formulated precise reporting standards for manuscripts submitted to APA journals, including reports of experimental evaluations of interventions and reports of meta-analyses (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008). The recommendations require that the statistical methods used to analyse data must be clearly described and justified. If

https://doi.org/10.1017/BrImp.2017.28 Published online by Cambridge University Press

TABLE 1

Recommendations of the Taskforce on Statistical Inference *APA Board of Scientific Affairs* Recommendations (Adapted from Wilkinson et al., 1999)

Power and sample size. *Provide information on sample size and the process that led to sample size decisions. Document the effect sizes, sampling and measurement assumptions, as well as analytic procedures used in power calculations. Because power computations are most meaningful when done before data are collected and examined, it is important to show how effect-size estimates have been derived from previous research and theory in order to dispel suspicions that they might have been taken from data used in the study or, even worse, constructed to justify a particular sample size. Once the study is analyzed, confidence intervals replace calculated power in describing results. (p 596)*

Hypothesis tests. *It is hard to imagine a situation in which a dichotomous accept-reject decision is better than reporting an actual p value or, better still, a confidence interval. Never use the unfortunate expression "accept the null hypothesis". Always provide some effect size estimate when reporting a p value.* (p 599)

Effect sizes. *Always present effect sizes for primary outcomes. If the units of measurement are meaningful on a practical level (e.g., number of cigarettes smoked per day), then we usually prefer an unstandardized measure (regression coefficient or mean difference) to a standardized measure (r or d). It helps to add brief comments that place these effect sizes in a practical and theoretical context. (p 599)*

Interval estimates. *Interval estimates should be given for any effect sizes involving principal outcomes. Provide intervals for correlations and other coefficients (p 599)*

inferential statistics are used, the *a priori* Type I error rate adopted must be reported. Differences between intended and actual sample size, power analysis (or methods used to determine precision of parameter estimates) should also be reported. It is not mandated that statistical significance levels (i.e., *p*-values) need to be reported, but findings must include ESs and CIs.

Prior to the publication of the 1999 APA-TSI recommendations, ESs were not frequently reported. In a survey of 391 articles that used inferential statistics published in four major psychology journals (three of which were APA journals), Kirk (1996) found that while just over 6% of articles reported three or more measures of ES, between 23% and 88% in the various journals did not report any such measure. Keselman et al. (1998) reviewed all articles using analysis of variance designs published in the 1994–1995 issues of 17 prominent education and psychology research journals. Only 41 (10%) of the 411 analyses examined reported ESs.

In a comprehensive review of 32 studies investigating ES reporting practices in psychology, education, behavioural research and rehabilitation journals, Peng, Chen, Chiang and Chiang (2013) found that since the publication of the 1999 APA-TSI recommendations reporting rate of ESs and, to a lesser extent, CIs had improved. Nonetheless, rate of reporting for ESs was still generally less than 50%, and even lower for CIs. A similar pattern was evident in the neuropsychological literature. In a survey of 406 articles published in *The Archives of Clinical Neuropsychology* between 1990 and 2004, Schatz, Jay, McComb and McLaughlin (2005) found that there was a significant increase in ES reporting after the release of the APA-TSI recommendations. Prior to the APA-TSI report (1990–1992) 4.3% of articles reported ESs. In the period commensurate with and shortly after the APA-TSI report (1996–2000), 22.4% reported ES, and in the subsequent period (2001–2004) ES was reported in 29.4% of articles.

Effect Size

A significant *p*-value indicates the likelihood of obtaining the observed data if H_0 is true. The magnitude of *p* (e.g., <0.05 vs. <0.0001) does not reflect the size (or the importance) of the intervention effect (Gliner, Leech, & Morgan, 2002). Smaller *p*-values (e.g., <0.0001) are, however, often and erroneously interpreted to mean that the intervention effect is larger. Cohen (1990) points out that '*Because science is inevitably about magnitudes, it is not surprising how frequently p values are treated as surrogates for effect sizes'*. (p. 1309). Glasser (1999) illustrates one aspect of this fallacy by considering a hypothetical Randomised Control Trial to investigate a dietary intervention for lowering cholesterol. The intervention reduces cholesterol levels from 243 to 238 mg/100 ml in the experimental group but there is no change in the control group. Glasser suggests that most clinicians might not consider changing clinical practice on the basis of such trivial effect (only 5 mg/100 ml), but would probably attribute more importance to the study findings if *p* was <0.05.

Cohen is credited with introducing the concept of ES to psychological research and measures of ES have been available for at least 60 years (Huberty, 2002). In essence, an ES provides measure

https://doi.org/10.1017/BrImp.2017.28 Published online by Cambridge University Press

of the degree to which the observed data differ from the expectations postulated in H_0 (Cohen, 1994). An ES quantifies the size of a difference between two groups and provides point estimate of treatment effect, that is, how well an intervention works. ESs are independent of sample size, whereas *p*-values reflect both ES and sample size. ESs are an important tool in reporting and interpreting treatment effectiveness. ESs can be calculated using either 'raw' differences between, say, the means of the treatment and control group (absolute ES), or as standardised differences. The problem with 'raw' ESs is that if the scale used to measure the outcome variable is not readily interpretable (e.g., rating on anxiety on a 0–10 scale) the meaning of the ES magnitude is unclear. Moreover, the magnitude of 'raw' ESs is influenced by the scale used to measure the outcome variable which further complicates interpretation. Standardised ESs take into account group variability are 'metric free' and are analogous to *z*-scores. Moreover, because they are 'metric free' the results of different studies addressing the same issue but using different outcome measures can be directly compared, making it possible to conduct meta-analyses.

ESs can be classified into two main 'families': the d-family for measures of between-group differences and the r-family, for measures of association (Ellis, 2010). A selection of common ESs in each family is shown in Tables 2 and 3.

As Cohen (1988) points out, hypotheses about differences between two means are the most frequently tested in behavioural research. Consequently, Cohen's ***d*** (Cohen, 1988) is a common used ES in these circumstances, which is fortunate because it, along with Glass's Δ (Glass, McGaw, & Smith, 1981) and Hedges'***g*** (Hedges, 1981), is easy to calculate. An example, using hypothetical data, of how to calculate each of these ESs is shown in Box 1. Each of these three ES measures expresses the impact of the intervention in terms of standard deviations. The magnitude of the ESs for the hypothetical data calculated by the three different methods is very similar. This is not surprising, given that the numerator in each formula is the same and only slightly different values for the standard deviation (σ) are used in the denominator of each formula.

So, which of the ESs should be calculated? If σ for the intervention and control groups are exactly equal, then Cohen's *d* can be calculated and either σ can be used; this, however, is an unlikely scenario in the real world. If σ for the two groups are reasonably equal, then it can be assumed that they are estimates of the population from which the two groups have been drawn (Ellis, 2010). In this instance, it would be also be appropriate to pool the standard deviations using the formula shown in Box 1 and calculate Cohen's ***d***. The question is: how 'reasonably equal' must the standard deviations of the two groups be? Unfortunately, there is no precise criterion for determining this, and researchers must exercise their own judgement. If in doubt, or if the standard deviations of the two are clearly unequal, then the homogeneity of variance assumption is violated and pooling the standard deviations is not appropriate. One solution is to calculate Glass's delta, using the control group σ. The assumption behind this is that the control group σ will not have been influenced by the intervention and will therefore reflect the population σ more faithfully (Ellis, 2010). The robustness of this assumption increases as the size of the control group increases. If the size of the intervention and control groups differs (particularly if the difference is substantial), it is more appropriate to calculate Hedge's *g* because it uses the weighted pooled σ for the two groups which takes into account group size. Moreover, it might, arguably, be more appropriate in 'real life' situations where sample sizes and/or σ for each sample might not always be equal. Turner and Bernard (2006) recommend its use because Hedge's ***g*** also corrects for the small-sample bias in Cohen's ***d*** and converges to it in large samples.

Other ESs listed in Tables 2 and 3 are measures of intervention effect applicable when data analysis involves more than comparison of two means (although a two group, one-way ANOVA data is essentially the same as a *t*-test). The computation for η^2 (top of Table 3) is straight forward from Sums of Squares in the SPSS® output. The formula shown in Table 3 is suitable for main effects in fixed factor designs, but can also be used for univariate contrasts by substituting the $SS_{\text{between groups}}$ in the numerator for SS_{contrast} (Olenik & Algina, 2000). Indeed, these authors elaborate the application of $\boldsymbol{\eta}^2$ to ANCOVA as well as multifactorial, random factors, and mixed designs.

Fortunately, the other ES indices in Table 3 can be directly obtained from SPSS® output. But this is not always the case. Statistical packages commonly used in psychological research, such as SPSS® and SAS®, can calculate other ESs, but this often requires specific syntax that is not available from the drop-down menus. Meyer, McGrath and Rosenthal (2003) provide syntax patches for calculating Pearson's ***r*** and Cohen's ***d*** in SPSS® and SAS®, as well as syntax for converting one type of ES to another. Torchiano (2017) has recently published a package of R-code to calculate various ESs. A more convenient way to calculate ESs, however, is to browse the free on-line calculators listed in Table 5. These sites offer calculators

https://doi.org/10.1017/BrImp.2017.28 Published online by Cambridge University Press

https://doi.org/10.1017/BrImp.2017.28 Published online by Cambridge University Press

TABLE 2
d Family Effect Sizes – Measures of between Group Differences

	ES	What it measures	Formula	Terms in formula
DICHOTOMOUS VARIABLES	**OR** Odds ratio	Odds of a desired outcome in the intervention group relative to the odds of the same outcome in the control group	$OR = \frac{a \times d}{b \times c}$	*a*=Frequency of successful outcomes in intervention group *b*=Frequency of non-successful outcomes in intervention group *c*=Frequency of successful outcomes in control group *d*=Frequency of non-successful outcomes in control group
CONTINUOUS VARIABLES	***d*** Cohen's *d*	Uncorrected standardised difference between two group means based on the pooled standard deviation	$d = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{\text{pooled}}}$	$\overline{X}_1$= Mean of group1; $\overline{X}_2$= mean of group2; σ_{pooled}= pooled standard deviation of both groups
	Δ Glass's delta	Uncorrected standardised difference between two group means based on the control group standard deviation	$\Delta = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{\text{control}}}$	$\overline{X}_1$= Mean of group1; $\overline{X}_2$= mean of group2; σ_{control}= standard deviation of control group
	g Hedges' *g*	Corrected standardised difference between two group means based on the pooled, weighted standard deviation	$g = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{w/\text{pooled}}}$	$\overline{X}_1$= Mean of group1; $\overline{X}_2$= mean of group2; $\sigma_{w/\text{pooled}}$= weighted pooled standard deviation of both groups.

https://doi.org/10.1017/Brimp.2017.28 Published online by Cambridge University Press

TABLE 3

r Family Effect Sizes – Measures of Association

		ES	What it measures	SPSS® procedure/Computation
Proportion of variance	ANOVA	η^2 Eta squared	proportion of the total variance attributable to a main effect	SPSS®: Analyse ➔ Compare Means ➔ ANOVA From output, calculate η^2 using the formula: $\eta^2 = \frac{SS_{between\ groups}}{SS_{total}}$ (Richardson, 2011) $SS_{between\ groups}$ = Sums of Squares between groups (Effect); SS_{total} = Total Sums of Squares
	ANCOVA	η^2 Eta squared	proportion of the total variance attributable to a main effect	SPSS®: Analyse ➔ GLM ➔ Univariate ➔ Options ➔ select Estimates of effect size
	MANOVA	partial η^2 partial Eta squared	like η^2 but with the effects of other independent variables and interactions partialled out	SPSS®: Analyse ➔ GLM ➔ Multivariate ➔ Options ➔ select Estimate of Effect Size
	Regression	r^2 Coefficient of determination	used in bivariate regression analysis	SPSS®: Analyse ➔ Regression ➔ Linear ➔ Statistics ➔ R^2 is default output which is the equivalent of r^2 in bivariate regression
		R^2 R squared	(uncorrected) coefficient of multiple determination	SPSS®: Analyse ➔ Regression ➔ Linear ➔ Statistics ➔ R^2 and $_{adj}R^2$ are default output
		$_{adj}R^2$ Adjusted R^2	R^2 adjusted for sample size and the number of predictor variables	
Correlation	r Pearson *r*		product moment correlation coefficient used when both variables an interval/ratio scale	SPSS®: Analyse ➔ Correlate ➔ Bivariate ➔ then choose Pearson in Correlation Coefficient box
	ρ Spearman's rho		rank correlation coefficient used when both variables are ordinal scale	SPSS®: Analyse ➔ Correlate ➔ Bivariate ➔ then choose Spearman in Correlation Coefficient box
	r_{pb} Point-biserial correlation coefficient		used when one of the two variables is dichotomous	SPSS®: Analyse ➔ Correlate ➔ Bivariate ➔ then choose Pearson in Correlation Coefficient box (NB: r_{pb} is a special case of Pearson's r when one of the variables is dichotomous)
	Φ Phi coefficient		measure of strength of association between two dichotomous variables	SPSS®: Analyse ➔ Descriptive Statistics ➔ Crosstabs ➔ Statistics, then select Phi coefficient

Box 1: Consider a study to investigate the effect of a brief memory training intervention on the ability of persons with stroke to recall stories. The hypothetical result for the two groups after the intervention is given below.

Intervention Group:	$N_I = 25$	Number of participants in intervention group
	$\overline{X}_I = 17.6$	Mean Number of Story Items recalled
	$\sigma_I = 3.4$	Standard deviation of intervention group
Control Group:	$N_C = 23$	Number of participants in control group
	$\overline{X}_C = 15.2$	Mean Number of Story Items recalled
	$\sigma_C = 4.2$	Standard deviation of control group

Cohen's *d* (Cohen, 1988): If the standard deviation of the two groups are equal (i.e., $\sigma_C = \sigma_I$) then either can be used to calculate ***d***, using the formula:

$$d = \frac{\overline{X}_I - \overline{X}_C}{\sigma}$$

If $\sigma_C \neq \sigma_I$, as is the case for the hypothetical data, Cohen suggests that the pooled standard deviation (σ_{pooled}) is calculated using the formula:

$$\sigma_{\text{pooled}} = \sqrt{\frac{\sigma_C^2 + \sigma_I^2}{2}}$$

and σ_{pooled} is then used instead of σ in the formula above.

For the hypothetical data: $\sigma_{\text{pooled}} = \sqrt{\frac{4.2^2+3.4^2}{2}} = \sqrt{\frac{17.6+11.6}{2}} = \sqrt{\frac{29.2}{2}} = \sqrt{14.6} = 3.8$

Calculating ***d*** for the hypothetical: $\boldsymbol{d} = \frac{\overline{X}_I - \overline{X}_C}{\sigma_{\text{pooled}}} = \frac{17.6-15.2}{3.8} = \frac{2.4}{3.8} = .63$

Glass'delta (Glass et al., 1981): Which group mean (i.e., $\overline{X}_I$ or $\overline{X}_C$) is substituted for $\overline{X}_1$ and $\overline{X}_2$ in Glass' formula is decided by the same considerations outlined above.

Calculating of Δ for the hypothetical: $\Delta = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{\text{control}}} = \frac{17.6-15.2}{4.2} = \frac{2.4}{4.2} = .57$

Hedges' *g* (Hedges, 1981): To calculate this ES, the weighted, pooled standard deviation needs to be calculated first. This can be done using the formula:

$$\sigma_{w/\text{pooled}} = \sqrt{\frac{\sigma_1^2 (N_1 - 1) + \sigma_2^2 (N_2 - 1)}{N_1 + N_2 - 2}}$$

For the hypothetical data:

$$= \sqrt{\frac{\sigma_I^2(N_I-1)+\sigma_C^2(N_C-1)}{N_I+N_C-2}} = \sqrt{\frac{3.4^2(25-1)+4.2^2(23-1)}{25+23-2}} = \sqrt{\frac{277.44+388.08}{46}}$$

$$= \sqrt{\frac{665.52}{46}} = \sqrt{14.5} = 3.80$$

Calculating ***g*** for the hypothetical data: $\boldsymbol{g} = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{w/\text{pooled}}} = \frac{\overline{X}_I - \overline{X}_C}{\sigma_{w/\text{pooled}}} = \frac{17.6-15.2}{3.8} = \frac{4.8}{3.8} = .63$

to compute ESs, CIs for ESs, transform one ES to another, as well as many other useful statistical tools. Peng, Chen, Chiang and Chiang (2013) have compiled a list of internet resources (Table 5, pp. 203–204) for calculating ESs.

Cohen's 'rules of thumb' for interpreting the magnitude of most common ESs have been universally adopted (Table 4) for other ESs. According to Cohen (1962), a medium ES is visible to the naked eye. He acknowledged that the magnitude

https://doi.org/10.1017/BrImp.2017.28 Published online by Cambridge University Press

TABLE 4
Interpretation of Effect Size

	Small	medium	large	Source
OR: Odds ratio	1.5	3.5	9.0	Ferguson, 2009
	1.5	2.5	4.3	Wilson, 2011
Cohen's ***d***	.20	.50	.80	Cohen, 1988
Glass's delta	.20	.50	.80	Ellis, 2010
Hedges' ***g***	.20	.50	.80	Ellis, 2010
η^2: Eta squared	.01	.06	.14	Cohen, 1988
	.02	.13	.26	Bakeman, 2005
partial η^2: partial Eta squared	.01	.13	.26	Draper, 2016
r^2: Coefficient of determination	.010	.059	.138	Cohen, 1988
R^2: ***R*** squared	.02	.13	.26	Cohen, 1988
$_{adj}R^2$: Adjusted R^2	.10	.30	.50	Kraemer et al., 2003
r: Pearson's ***r***	.10	.30	.50	Cohen, 1988
	.10	.25	.40	Wilson, 2011
ρ: Spearman's rho	0 - .3			
r_{pb}: Point-biserial correlation coefficient	.10	.24	.37	Kirk, 1996
Φ: Phi coefficient	< .20	.20 - .60	> .6	Rea & Parker, 1992
	.10	.30	.50	Cohen, 1988

of what he labelled small, medium and large ESs was chosen arbitrarily and cautioned the reader to exercise their own judgement. Opinion differs on how to interpret magnitude of other ESs. For example, Kraemer et al. (2003) classify the magnitude of d-family ESs as: much larger than typical >1.0, large of larger than typical = .80, medium or typical = .50 and small or smaller than typical = .20. Similarly, they classify the magnitude of r-family effects as: much larger than typical >.70, large or larger than typical = .50, medium or typical = .30 and small or smaller than typical =.10.

Confidence Intervals

The CI of the ES provides the precision of the estimation. CIs are, in essence, error terms that provides the likely range of the intervention effect (Fethney, 2010; Turner & Bernard, 2006). As power increases with larger sample sizes the CI range diminishes, indicating greater ES precision (Halsey, Curran-Everett, Vowler, & Drummond, 2015). Estimating CIs for some ESs can be relatively straight forward. Berben, Sereika and Engberg (2012) describe a simple method for calculating the 95% CI for Cohen's ***d***. The method is illustrated in Box 2, using the hypothetical data from Box 1. In this instance, ES = .63 and the 95% CI = .04–1.22. What this CI tells is that true ES in the population from which the two hypothetical groups (i.e., N_I and N_C) were drawn is 95% certain to be in the range of .04–1.22. Because the CI does not include zero, we can reasonably assume that there is a positive intervention effect. If the CI had included zero, it would raise the possibility that the observed ES might have been obtained by chance.

Although formulae for calculating CI are available (e.g., Cumming & Finch, 2001; Li-Ting & Chao-Ying, 2013), they either require computer intensive iteration procedures or the hand calculations are cumbersome. However, one 'complication' with the calculation of ES CIs is that non-central *t*-distributions (i.e., an asymmetrical distribution not distributed around zero) are used (Howell, 2010). Smithson (2001) provides SPSS® syntax to estimate non-central CIs for various ESs. Fortunately, however, most of the on-line calculators listed in Table 5 also compute CIs.

Closing Remarks

NHST does not do what we think it does and the way it works is often misunderstood, but it is too entrenched to disappear completely any time soon. For almost 20 years, there have been clear directives that researches should not rely solely on NHST and report only *p*-values. A statistically significant result (i.e., $p < 0.05$) allows the researcher to decide with some degree of confidence that observed differences between groups, namely the ES, are reliable if H_0 were true. This is particularly useful if sample sizes are small. However, a result can be statistically significant merely because the sample size is large, even if the ES is trivial. Moreover, statistical significance is not equivalent

https://doi.org/10.1017/BrImp.2017.28 Published online by Cambridge University Press

Box 2: Calculation of Confidence Interval for Cohen's ***d***

The first step is to calculate the standard deviation of ***d*** using the formula:

$$\sigma_d = \sqrt{\frac{N_I+N_C}{N_I \times N_C} + \frac{d^2}{(2/N_I+N_C)}}$$ (Cooper, Hedges, & Valentine, 2009)

Using the hypothetical data from Box 1: $N_I = 25$, $N_C = 23$ and ***d*** =.63

$$\begin{aligned}\sigma_d &= \sqrt{\frac{25+23}{25\times23} + \frac{.63^2}{(2/25+23)}} \\ &= \sqrt{\frac{48}{575} + \frac{.3969}{48}} \\ &= \sqrt{0.083 + 0.008} = 0.30\end{aligned}$$

The 95% CI is calculated using the formula:

95% CI $= d \pm (z_{0.05/2} \times \sigma_d)$ where $z_{0.05/2}$ is the z-score for a two-tailed probability of 0.05, which is 1.96. Hence:

$$\begin{aligned}\text{95\% CI} &= .63 \pm (1.96 \times 0.30) \\ &= .63 \pm (0.59)\end{aligned}$$

Which means that ***d*** ranges between 0.04 and 1.22

TABLE 5

Free On-line Calculators

Effect Size Calculators: https://www.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html

These calculators were developed by Paul D Ellis of the Hong Kong Polytechnic University. Ellis is the author a comprehensive book on effect sizes: Ellis PD (2010) The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results. New York, Cambridge University Press. Cohen' ***d,*** Hedges' ***g*** and Glass' Δ can be computed using either means and standard deviations, t-statistic and sample size, or the correlation coefficient *r*. There are also calculators to compute strength of association (*r*), using either d (equal or unequal groups), Chi square with df=1, or *z* scores.

Free Statistics Calculators, Version 4.0: http://www.danielsoper.com/statcalc/related.aspx?id=96

This site was created by Daniel Sloper, Associate Professor in the Department of Information Systems and Decision Sciences, Mihaylo College of Business and Economics, California State University, Fullerton. It contains 106 free statistics calculators, including several to calculate ESs and CIs for ESs and Power.

Effect Size Psychometrica: https://www.psychometrica.de/effect_size.html

Has several online calculators for the computation of different common ESs, for nonparametric tests (Mann–Whitney-U, Wilcoxon-W and Kruskal–Wallis-H). It also has calculators to transform one ES to another (e.g., d to η^2), and provides a table for interpreting ES magnitude for *d*, *r* and η^2.

Practical Meta-Analysis Effect Size Calculator: https://www.campbellcollaboration.org/escalc/html/EffectSizeCalculator-Home.php

This calculator by David B Wilson of George Mason University, is a companion to the book Lipsey MW and Wilson DB (2001) Practical Meta-Analysis. Sage Publications, Thousand Oaks. It computes ESs and CIs for standardised mean differences, the correlation coefficient, odds ratios and risk ratios.

to clinical significance, that is whether or not an intervention produces a change that is important and useful in the real world (see for example, Fethney, 2010; Sainani, 2012; Thompson, 2002). On the other hand, even small ESs might be clinically significant. Prentice and Miller (1992) suggest that a small ESs might be important when either a small change in the independent variable (for our purposes, the intervention) produces a change in the dependent variable (for our purposes, the target behaviour or primary outcome measure), or when a change is elicited on a dependent variable that is resistant to change.

P-values give us useful information about the probability that the obtained data pertains to samples drawn from the same or different populations. What we want to know in neurorehabilitation is whether or not an intervention has had a 'real' or clinically significant effect: measures of ES and their CIs can give us that information.

https://doi.org/10.1017/BrImp.2017.28 Published online by Cambridge University Press

Financial Support

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Conflict of Interest

None.

Ethical Standards

The research does not involve human experimentation.

References

APA Publications and Communications Board Working Group on Journal Article Reporting Standards (2008). Reporting standards for research in psychology: Why do we need them? What might they be?. *American Psychologist, 63*(9), 839–851.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*(3), 379–384.

Berben, L., Sereika, S.M., & Engberg, S. (2012). Effect size estimation: Methods and examples. *International Journal of Nursing Studies, 49*, 1039–1047.

Berkson, J. (1938). Some difficulties of interpretation encountered in application of Chi squared. *Journal of the American Statistical Association, 33*(203), 526–536.

Carver, R.P. (1978). The case against statistical significance. *Harvard Educational Review, 48*(3), 378–399.

Castro Sotos, A.E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*(2), 98–113.

Clark, C.A. (1963). Hypothesis testing in relation to statistical methodology. *Review of Educational Research, 33*, 455–473.

Cohen, J. (1962). The statistical power of abnormal–social psychological research. *Journal of Abnormal and Social Psychology, 65*, 145–153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (So far). *American Psychologist, 45*(12), 1304–1312.

Cohen, J. (1994). The Earth is round ($p < .5$). *American Psychologist, 49*(12), 997–1003.

Cooper, H., Hedges, L.V., & Valentine, J.C. (2009). *The handbook of research and synthesis and meta-analysis*. New York: Russell Sage Foundation.

Cumming, G. & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*(4), 532–574.

Draper, S.W. (2016). *Effect Size*. Retrieved from http://www.psy.gla.ac.uk/~steve/best/effect.html

Ellis, P.D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. New York: Cambridge University Press.

Falk, R. & Greenbaum, C.W. (1995). Significance tests die hard. The amazing persistence of a probabilistic misconception. *Theory and Psychology, 5*(1), 76–98.

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

Fethney, J. (2010). Statistical and clinical significance, and how to use confidence intervals to help interpret both. *Australian Critical Care, 23*, 93–97.

Fisher, R.A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences. Methodological issues* (pp. 311–339). Hillsdale, NJ: Lawrence Erlbaum Associates.

Glaser, D.N. (1999). The controversy of significance testing: Misconceptions and alternatives. *American Journal of Critical care, 8*(5), 291–296.

Glass, G.V., McGaw, B., & Smith, M.L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage

Gliner, J.A., Leech, N.L., & Morgan, G.A. (2002). Problems with null hypothesis significance testing (NHST): What do the textbooks say?. *The Journal of Experimental Education, 71*(1), 83–92.

Halsey, L.G., Curran-Everett, D., Vowler, S.L., & Drummond, G.B. (2015). The fickle P value generates irreproducible results. *Nature Methods, 12*(3), 179–185.

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 106–128.

Howell, D.C. (2010). Confidence intervals on effect size. Retrieved from: https://www.uvm.edu/~dhowell/methods7/Supplements/Confidence%20Intervals%20on%20Effect%20Size.pdf

Huberty, C.J. (2002). A history of effect size indices. *Educational and Psychological Measurement, 62*(2), 227–240.

Huberty, C.J., & Pike, C.J. (1999). On some history regarding statistical testing. *Advances in Social Science Methodology, 5*, 1–22.

Keselman, H.J., Huberty, C.J., Lix, L.M., Olejnik, S., Cribbie, R., Donahue, B., . . . Levin, J.R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*, 350–386.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*(5), 746–759.

https://doi.org/10.1017/BrImp.2017.28 Published online by Cambridge University Press

Kraemer, H.C., Morgan, G.A., Leech, N.L., Gliner, J.A., Vaske, J.J., & Harmon, R.J. (2003). Measures of clinical significance. *Journal of the American Academy of Child and Adolescent Psychiatry, 42*(12), 1524–1529.

Krishnan, S. & Idris, N. (2014). Students' misconceptions about hypothesis test. *REDIMAT: Journal of Research in Mathematics Education, 3*(3), 276–293.

Lambdin, C. (2012). Significance tests as sorcery: Science is empirical-significance tests are not. *Theory and Psychology, 22*(1), 67–90.

Li-Ting, C., & Chao-Ying, J.P. (2013). Constructing confidence intervals for effect sizes in ANOVA designs. *Journal of Modern Applied Statistical Methods, 12*(2), 82–104.

Meehl, P.E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*(2), 103–115.

Meyer, G.J., McGrath, R.E., & Rosenthal, R. (2003*). Basic effect size guide with SPSS® and SAS® syntax*. Retrieved from www.tandf.co.uk/journals/authors/hjpa/resources/basiceffectsizeguide.rtf.

Neyman, J., & Pearson, E. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, *20A*, 175–240.

Neyman, J., & Pearson, E. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika*, *20A*, 263–294.

Nickerson, R.S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241–301.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25*, 241–286.

Peng, C.-Y. J., Chen, L.-T., Chiang, H.-M., & Chiang, Y.-C. (2013). The impact of APA and AERA guidelines on effect size reporting. *Educational Psychology Review, 25*, 157–209.

Prentice, D.A., & Miller, D.T. (1992). When small effects are impressive. *Psychological Bulletin, 112*(1), 160–164.

Rea, L.M., & Parker, R.A. (1992). *Designing and conducting survey research*. San Francisco: Jossey-Boss.

Richardson, J.T.E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review, 6*(12), 135–147.

Rozeboom, W.W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57*(5), 416–428.

Sainani, K.L. (2012). Clinical versus statistical significance. *American Academy of Physical Medicine and Rehabilitation, 4*(6), 442–445.

Schatz, P., Jay, K.A., McComb, J., & McLaughlin, J.R. (2005). Misuse of statistical tests in *archives of clinical neuropsychology* publications. *Archives of Clinical Neuropsychology, 20*, 1053–1059.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement, 61*(4), 605–632.

Thompson, B. (2002). "Statistical," "Practical," and "Clinical": How many kinds of significance do counselors need to consider?. *Journal of Counselling and Development, 80*, 64–71.

Torciano, M. (2017*) Efficient effect size computation*. Retrieved from https://cran.r-project.org/web/packages/effsize/effsize.pdf

Turner, H.M., & Bernard, R.M. (2006). Calculating and synthesizing effect sizes. *Contemporary Issues in Communication Science and Disorders, 33*, 42–55.

Vallecillos, A. (2001). Cuestiones metodológicas en la investigación educativa. Quinto Simposio de la Sociedad Española de Investigación en Educación Matemática, Almería, Spain.

Vallecillos, A., & Batanero, C. (1997b). Conceptos activados en el contraste de hipótesis estadísticas y su comprensión por estudiantes universitarios. *Recherches en Didactique des Mathématiques, 17*(1), 29–48.

Vallecillos, A., & Batanero, M.C. (1997a). Aprendizaje y enseñanza del contraste de hipotesis: Concepciones y errores. *Enseñanza de las Ciencias, 15*(2), 189–197.

Wilkinson, L. and the Task Force on Statistical Inference APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604.

Wilson, D.B. (2011). *Interpretation.ppt*. Retrieved from http://mason.gmu.edu/~dwilsonb/ma.html.

https://doi.org/10.1017/BrImp.2017.28 Published online by Cambridge University Press