Pergamon

Archives of Clinical Neuropsychology
16 (2001) 653–667

Archives
of
CLINICAL
NEUROPSYCHOLOGY

Statistics to tell the truth, the whole truth, and nothing but the truth: formulae, illustrative numerical examples, and heuristic interpretation of effect size analyses for neuropsychological researchers

Konstantine K. Zakzanis*

Division of Life Sciences, University of Toronto, 1265 Military Trail, Toronto, Ontario, Canada M1C 1A4

Accepted 10 July 2000

Abstract

If, as neuropsychologists, we think of the relationship between brain and behavior as the same as that between truth and reality, we must be equipped with statistical procedures that are coherent in terms of what we measure and what it represents. I believe that this necessary statistical procedure is effect size analysis, and without it, I believe that we fail to tell the truth, the whole truth, and nothing but the truth when describing our neuropsychological research. Accordingly, I review here the standard calculations of commonly employed effect sizes in two group designs and show how to adjust some familiar (and perhaps not so familiar) formulae using illustrative numerical examples. I also put forth an argument to adopt Cohen's measure as an expression of effect size based on its apropos to neuropsychological research. It is also argued that the interpretation of the magnitude of an effect size should depend on context, and not on pre-established heuristic benchmarks. It is noted, however, that effect sizes greater than 3.0 (OL% < 5) might seem particularly appropriate when evaluating the sensitivity of neuropsychological tasks and in establishing test markers in neuropsychological disorders. © 2001 National Academy of Neuropsychology. Published by Elsevier Science Ltd.

Keywords: Neuropsychology; Statistical analysis; Effect size formulae; Methodology

In the Tractatus Logico-Phiosophicus, Wittgenstein (1949) developed a theory of truth. He noted that basic statements, concerning, as it were, the atoms of knowledge and experience, directly correspond to reality. From them, other more complex statements were derived, the

* Tel.: +1-416-287-7424; fax: +1-416-287-7642.
E-mail address: zakzanis@scar.utoronto.ca (K.K. Zakzanis).

0887-6177/01/$ – see front matter
PII: S0887-6177(00)00076-7

truth of which depended on their consistency or coherence with the constituent basic statements. As far as the basic statements were concerned, Wittgenstein thought of the relationship between truth and reality as the same as that between a picture and what it represents.

Now, accept that in neuropsychological research, our atoms of knowledge are really just data and our experience is our observation of how brain is related to behavior. To make our experience a reality, we employ more complex statements regarding our observations by way of statistical analysis. The truth of which should depend on their consistency or coherence with the constituent basic statements. But do they? As far as the basic statements are concerned, neuropsychologists think of the relationship between brain and behavior as the same as that between truth and reality. If we measure our reality (i.e., data) with statistical procedures that are incoherent, then are we not marring our truth and reality regarding brain–behavior relationships?

What is this incoherent statistical procedure that I speak of? It is null hypothesis statistical significance testing (NHSST). I do not mean to equate NHSST with the devil himself, as appropriate use of these methods do indeed exist (see Frick, 1996). What I do mean to equate NHSST is with its limitation in neuropsychological research; although this is not the main impetus for this work. Several other works exist that demonstrate vividly the illogical statistical intricacies of NHSST in psychological (e.g., Bakan, 1966; Carver, 1978; Cohen, 1994; Guttman, 1985; Harlow, Mulaik, and Steiger, 1997; Kirk, 1996; Rozeboom, 1960; Schmidt, 1996; Snyder & Lawson, 1993; Thompson, 1998; Thompson & Snyder, 1998; for a forthcoming thorough review of effect sizes, see Olejnik & Algina, in press) and neuropsychological research (Bieliauskas, Fatenau, Lacy, & Roper, 1997; Rourke, Costa, Cicchetti, Adams, & Plasterk, 1991; Soper, Cichetti, Satz, Light, & Orsini, 1988; Zakzanis, 1998). Most of these papers also argue for the due consideration of effect sizes as complementary statistics in research reports if, as scientists, we are really interested in estimating the magnitude of the parameters in the populations and in recognizing the phenomenon that may be manifesting themselves in these magnitudes. The purpose of this work is to provide neuropsychological researchers with statistical procedures that are more likely to give truth to our reality of brain–behavior relationships. Statistics that will more readily allow the neuropsychologist to estimate and recognize the magnitude of relation between brain and behavior. That is, I present formulae, illustrative numerical examples, and heuristic interpretation of effect size analysis that might aid further the neuropsychological researcher in their quest to understand brain and behavior.

Finally, the reader should be alerted that I would argue for always reporting effect sizes. Although this opinion has been voiced for almost 40 years, researchers have seldom reported effect sizes in their works despite APA Task Force Reports (e.g., Wilkinson & The APA Task Force on Statistical Inference, 1999). For example, the APA Task Force on Statistical Inference recently recommended to

> '*Always* provide some effect-size estimate when reporting a *P* value' (p. 599, emphasis added), and noted that 'We must stress again that reporting and interpreting effect sizes in the context of previously reported effects is *essential* to good research' (Wilkinson & The APA Task Force on Statistical Inference, 1999, p. 599, emphasis added). It is also noted, 'Unfortunately, ... the effect size of the APA publication manual encouragement has been negligible' (p. 599).

That is, there are now 11 empirical studies of one or two volumes of 23 different journals (Vacha-Haase, Nilsson, Reetz, Lance, & Thompson, 2000) showing no appreciable increase in effect size reporting following the 1994 APA publication manual "encouragement" (p. 18). This utter dearth of reporting effect sizes might mean that an APA "encouragement" is too vague to be enforceable. Moreover, only "encouraging" effect size reporting

> . . . presents a self-canceling mixed message. To present an 'encouragement' in the context of strict absolute standards regarding the esoterics of author note placement, pagination, and margins is to send the message, 'these myriad requirements count, and this encouragement doesn't' (Thompson, 1999b, p. 162).

Accordingly, it would seem just to use this opportunity to send a message to the editor, editorial board, and readers of Archives of Clinical Neuropsychology whom I suspect are mostly members of the National Academy of Neuropsychology. That is, should the journal require that effect sizes be always reported in research studies submitted to the journal? There are 11 journals that already do (e.g., Educational and Psychological Measurement, Journal of Applied Psychology, Journal of Early Intervention, Journal of Experimental Education, and Journal of Learning Disabilities) with more on the way.

It has been suggested that "there is only one force that can effect a change, and that is the same force that helped institutionalize null hypothesis testing as the sine qua non for publication, namely, the editors of the major journals" (Sedlmeier & Gigerenzer, 1989, p. 315). Hence, I suspect the message to follow below should solicit commentary from editorial board members, readers of the journal, and Academy members aimed in a constructive manner at the editor of this journal.

1. Statistics to tell the truth, the whole truth, and nothing but the truth

The Publication Manual of the American Psychological Association (American Psychiatric Association, 1994) "now 'encourages' researchers" to provide and interpret effect size information, yet, even the most cursory glance at leading research journals in neuropsychology, however, will reveal that authors seldom report effect sizes. One plausible explanation for this neglect may be that few researchers have a clear idea of when or how to calculate and interpret them. Hence, I will begin by reviewing the standard calculations of commonly employed effect sizes in two group designs and show how to adjust some familiar (and perhaps not so familiar) formulae.

In a two-group design, when we are interested in discovering the magnitude of difference between two sample means, or the departure from the null hypothesis in standardized units for fixed or random effects models, we can employ either Eq. (1) or Eq. (2):

$$\text{Hedge's } g = \frac{[M_1 - M_2]}{\text{S.D.}_{\text{control}}} \tag{1}$$

$$\text{Cohen's } d = \frac{[M_1 - M_2]}{\text{S.D.}_{\text{pooled}}} \tag{2}$$

When the sample N is equal between groups, Cohen's d requires the computation of a pooled standard deviation (S.D.) taking the form of Eq. (3):

$$\text{S.D.}_{\text{pooled}} = \frac{\text{S.D.}_1 + \text{S.D.}_2}{2} \tag{3}$$

When N is not equal between groups, however, a pooled S.D. weighted by sample size needs to be calculated to obtain Cohen's d using Eq. (4):

$$\text{S.D.}_{\text{pooled}} = \sqrt{\frac{(N_1 - 1)\text{S.D.}_1^2 + (N_2 - 1)\text{S.D.}_2^2}{N_1 + N_2 - 2}} \tag{4}$$

Note that effect sizes (whether they be Hedges' or Cohen's measure) must always be computed using the S.D. and not the standard error of the mean (S.E.M.). The consequence of computing an effect size using the S.E.M., is an over-inflated estimate of effect size. Hence, to transform the S.E.M. to S.D. (Eq. (5)):

$$\text{S.E.M. to S.D.} = (\text{S.E.M.})(\sqrt{N}) \tag{5}$$

If one is uncertain which of the two measures of effect size is most appropriate (I will, however, propose to adopt Cohen's d as a more appropriate measure for research in neuropsychology below), we can also transform g to d and d to g if desired using Eqs. (6,7):

$$g \text{ to } d = \sqrt[g]{\frac{df_{\text{within}}}{N}} \tag{6}$$

$$d \text{ to } g = \sqrt[d]{\frac{N}{df_{\text{within}}}} \tag{7}$$

If one is most familiar with the vagaries of correlational analysis, we can similarly transform Hedges' g into r by Eq. (8):

$$r = \frac{g}{\sqrt{g^2 + 4\left[\frac{df_{within}}{N}\right]}} \tag{8}$$

or Cohen's d into r by Eq. 9:

$$r = \frac{d}{\sqrt{d^2 + 4}} \tag{9}$$

This would then allow the neuropsychological researcher to interpret the magnitude of their obtained effect size by way of correlational magnitude. Of course, one could then square the obtained r to obtain a coefficient of determination or, less formally, the proportion of shared variance (R^2).

If our research design is to employ a fixed effects model, we can also use Ω^2 as a measure of effect size (Eq. (10)):

$$\Omega^2 = \frac{t^2 - 1}{t^2 + df + 1} \tag{10}$$

Without elaborating the use of Ω^2 further, however, it is at this point that I shall simplify the preceding array of choices for expressing our calculation of effect size. That is, if we are to adopt the effect size measure into our neuropsychological research, it seems wise to settle upon the best expression. I will propose the use of Cohen's *d* as a preferred expression of effect size by the process of elimination. It is important however, to keep in mind that Cohen's *d* is not meant to replace another effect size index where one might be more useful in a different situation. Accordingly, although I will champion the use of Cohen's *d* in neuropsychological research, keep in mind that it is only one formula of dozens of available choices.

Firstly, although Ω^2 is a "corrected" variance-accounted-for effect size (see Snyder & Lawson, 1993) where part of this correction is to adjust for positive bias associated with smaller sample sizes, the statistic incorporates a *t* value into its formulation. If we accept that a *t* statistic is biased in keeping with the observation that the value for *t* can be manipulated directly by total sample size (see Bakan, 1966; Meehl, 1967), and in keeping with the rationale for the inclusion of effect sizes into our research reports (see Schmidt, 1996; Zakzanis, 1998), it would seem hypocritical, if not deceiving to argue for the use of Ω^2 as a logically valid expression of effect size. The same argument can be made regarding the use of R^2 without much elaboration. That is, Meehl (1967) convincingly showed the waggish result of using an extremely large *N* in correlational analysis. Again, it would seem inappropriate to use a fallacious statistic to "tell the truth." This leaves us with Hedges' *g* and Cohen's *d*. Neither can be used without accepting their inherent biases however. Hence, we should be interested in putting to work the measure that is most appropriate for neuropsychological research, and it is in this distinction that Cohen's *d* is argued to be the most appropriate statistic to "tell the truth." That is, although some prefer Hedges' measure because its calculation is elementary, and its bias can be corrected by the use of the expected value of the reciprocal of a sample S.D., Cohen's *d* accounts explicitly for the variability typically seen in patients with neuropsychological disorders. That is, Hedges' *g* is based on the difference between patient and control means calibrated by the control S.D. only. There seems to be an inherent assumption in the computation of Hedges' measure that the patient S.D. is not much different from the S.D. of the control sample. What if the patient sample S.D. was much different? If it were considerably larger, then Hedges' *g* would certainly overestimate the magnitude of the effect size. Conversely, if the S.D. of the patient sample were considerably smaller, then Hedges' *g* would underestimate the magnitude of the effect size. Because Cohen's *d* is the difference between patient and control means calibrated in pooled S.D. units, the assumption of homogeneity of variance need not be assumed. Moreover, it is atypical in neuropsychological research to find proportional S.D.'s between patient and control samples (e.g., see Valdois, Joanette, Poissant, Ska, & Dehaut, 1990). Hence, although Cohen's *d* requires the additional labors of computing a pooled S.D., it is nonetheless a straightforward calculation (see Eqs. (3) and (4)), and more importantly, it does not fall victim to assumption of variance homogeneity. Moreover, Cohen (1988) provides estimates of distribution

nonoverlap associated with most values of *d*. With a simple conversion, these estimates can be presented in terms of overlap. That is, by subtracting the nonoverlap from 100, Table 1 provides overlap values for Cohen's *d* ranging from 0.0 (complete overlap) to 4.0 (relative no overlap). The overlap statistic (OL%) is meant to represent the amount of test measure overlap between patient and control samples. That is, OL% reflects the amount of overlap in the distribution of test measure scores between patient and control samples. For example, if a researcher were to administer the Boston Naming Test (Kaplan, Goodglass, & Weintraub, 1983) to a sample of patients with Parkinson's disease (PD) and healthy controls and found an effect size of 0.1 corresponding to 92.3% overlap, this would mean that only 7.7% of the patients with PD obtained scores that were not obtained by the healthy controls. In other

Table 1
Overlap percentages for values of *d*

d	Percent overlap	Interpretation
0.0	100	No effect
0.1	92.3	
0.2	85.3	
0.3	78.7	
0.4	72.6	
0.5	66.6	
0.6	61.8	
0.7	57.0	
0.8	52.6	
0.9	48.4	
1.0	44.6	
1.1	41.1	
1.2	37.8	
1.3	34.7	
1.4	31.9	
1.5	29.3	
1.6	26.9	
1.7	24.6	
1.8	22.6	
1.9	20.6	
2.0	18.9	
2.2	15.7	
2.4	13.0	
2.6	10.7	
2.8	8.8	
3.0	7.2	Clinical marker criteria
3.2	5.8	
3.4	4.7	
3.6	3.7	
3.8	3.0	
4.0	2.3	Approximate absolute discriminability

Based on Cohen's (1988) idealized population distribution.

Cohen (1988) presents "nonoverlap" values associated with *d*. Table 3 presented above displays "overlap" values associated with *d*, which are simply subtracted from the nonoverlap values adding, of course, to 100.

words, about 92% of the patient scores fell within the distribution of scores obtained by the healthy controls. Conversely, if a different researcher were to administer the Boston Naming Test to a sample of patients with probable Alzheimer's disease (AD) and healthy controls and found an effect size of 3.0 corresponding roughly to 7% overlap, this would indicate that the test was reliably sensitive to the presence of AD. That is, 93% of the patients obtained scores on the Boston Naming Test unlike any of those obtained by the healthy controls. Moreover, Table 1 also includes comments on the heuristic interpretation of the size of effect. I will return to this later in the section incorporating the illustrative numerical examples.

With these considerations in mind, it would seem most appropriate then that Cohen's *d* is best suited to express effect size in neuropsychological research.

If one is engaged in the review of existing literature and is at the same time interested in gauging the magnitude of effect in terms of Cohen's *d*, one will quickly discover just how seldom effect sizes are reported, let alone Cohen's *d*. Fortunately, formulae do exist, which can aid the researcher in transforming commonly reported test statistics into *d*. These formulae are presented in Tables 2 and 3. Table 2 includes those conversion formulae to obtain *r*. Table 3 includes those conversion formulae to obtain *d*. I shall walk through some illustrative numerical examples shortly.

At some point in the process of reviewing a literature, one might be tempted to conduct a more formal meta-analysis. Many excellent texts and papers have been written that detail such procedures (e.g., Cooper & Hedges, 1994; Hedges & Olkin, 1985; Rosenthal, 1991, 1995; Wolf, 1986), and I will not provide such a review here. There is, however, a statistic that meta-analysts commonly employ that will indeed help us discover the truth in our data analysis, which I would like to detail. That is, when reviewing a literature to determine the average effect size across published studies, it is unlikely that such a review will uncover every study that meets one's inclusion criteria. Rosenthal (1979) has called this the "file drawer problem" because of the tendency for studies supporting the null hypothesis of no significant results to be more likely to be buried away in file drawers. Kraemer and Andrews (1982) note that "published research studies tend to be biased toward positive findings. A study is often abandoned if it is apparent that statistically significant findings will not be forthcoming. Reports of nonsignificant findings are generally unpublishable even when they are replications of earlier studies reporting significant results" (p. 405). As Wolf (1986) adds, "even if

Table 2
Guidelines for converting various test statistics to r

Statistic to be converted	Formula for transformation to *r*	Comment
t	$r=\sqrt{\frac{t^2}{t^2+df}}$	
F	$r=\sqrt{\frac{F}{F+df\,(\text{error})}}$	Use only for the comparison of two group means (i.e., numerator $df=1$)
X^2	$r=\sqrt{\frac{X^2}{n}}$	Use only for 2×2 frequency tables ($df=1$); $n=$ sample size
d	$r=\frac{d}{\sqrt{d^2+4}}$	

Taken from Wolf (1986).

Table 3
Guidelines for converting various test statistics to *d*

Statistic to be converted	Formula for transformation to *d*	Comment
t	$d=\frac{2t}{\sqrt{df}}$	
F	$d=\frac{2\sqrt{F}}{\sqrt{df\,(\text{error})}}$	Use only for the comparison of two group means (i.e., numerator $df=1$)
r	$d=\frac{2r}{\sqrt{1-r^2}}$	

Taken from Wolf (1986).

investigators submit results of these studies for publication, it is generally difficult for editors to publish them because of the many studies they receive with statistically significant results. This may enhance the likelihood of a type I publication bias error in finding more positive results than is really the case were all studies to be located and included in our review'' (p. 38).

Rosenthal (1979) suggested that we can partly address this issue by calculating the number of studies confirming the null hypothesis that would be needed to reverse a conclusion that a significant relationship exists. Cooper (1979) called this the Fail Safe *N* (N_{fs}) for the number of additional studies in a meta-analysis that would be necessary to reverse the overall probability obtained from our combined test to a value higher than our critical value for statistical significance, usually 0.05 or 0.01 (Wolf, 1986).

Orwin (1983) has provided a fail-safe *N* formula (Eq. (11)):

$$N_{\text{fs}} = \frac{N(d - d_{\text{c}})}{d_{\text{c}}} \tag{11}$$

where N = the number of studies in the meta-analysis, d = the average effect size for the studies synthesized, and d_{c} = the criterion value selected that d would equal when some knowable number of hypothetical studies (N_{fs}) were added to the meta-analysis. Because of the lack of formally agreed upon criterion d values for effect size, Orwin (1983) suggests using Cohen's (1988) suggestion of $d=0.2$ (small effect) for d_{c}. Hence, N_{fs} will estimate the number of additional hypothetical studies needed that will overturn your obtained mean effect size to a small, typically meaningless effect size of 0.2.

2. Illustrative numerical examples

Table 4 displays hypothetical Boston Naming Test raw scores for 10 patients with probable AD and 10 normal controls. The duration of illness for each of the patients is also included along side the raw scores. Tests of significance were carried out using *t* and *F*. We can see that both tests were significant at the .001 level. Hence, we can conclude that our findings were unlikely to be chance. We can also say that there is a difference between the sample means; the null hypothesis of no difference cannot be accepted. Can we say much else in terms of statistical significance? The answer is no if we wish to tell the truth. That is, the truth of our analysis of the data should depend on their coherence with the constituent basic statements. In

Table 4
Hypothetical Boston Naming Test performance in patients with probable dementia of the Alzheimer's Type (AD) and Normal Controls (NC)

	AD ($N=10$)	Duration of illness (years)	NC ($N=10$)
Raw scores	20	5	56
	50	1	51
	18	6	60
	33	3	58
	39	3	53
	40	4	55
	16	7	54
	47	2	51
	55	1	59
	52	1	58
Mean (S.D.) (S.E.)	37.0 (14.7) (4.6)		55.5 (3.2) (1.0)

t test (18) = − 3.89, $P<.001$
F test (1, 18) = 15.16, $P<.001$
Pearson correlation for duration of illness and raw score on the Boston Naming Test in patients with DAT: $r=-9.6$, $P<.001$

other words, if we wish to describe how large the difference in sample means is, we need to employ those statistics that can tell us so. If we merely rely on our computed P value to tell us the truth about our finding, we will not reach a coherent conclusion. That is, Thompson (1999a) noted that,

> The calculated P values in a given study are a function of several study features, but are particularly influenced by the confounded, joint influence of study sample size and study effect sizes. Because P values are confounded indices, in theory 100 studies with varying sample sizes and 100 different effect sizes could each have the same single $P_{\text{calculated}}$, and 100 studies with the same single effect size could each have 100 different values for $P_{\text{calculated}}$. (pp. 167–168).

Thus, to obtain Hedges' measure of effect, we employ Eq. (1),

$$\text{Hedge's } g = \frac{37.0 - 55.5}{3.2} = -5.78 \ (\text{OL\%} < 2)$$

To obtain Cohen's d, which in this example is a much fairer estimate of effect given the heterogeneity of variance, we employ Eqs. (2) and (3),

$$\text{Cohen's } d = \frac{37.0 - 55.5}{14.7 + 3.2/2} = -2.07 \ (\text{OL\%} = 18)$$

Let us suppose for a moment that there were unequal sample sizes. Say 15 patients with AD and 10 normal controls. To obtain Cohen's d, we would then employ Eq. (4) to obtain the pooled S.D.

$$\text{S.D.}_{\text{pooled}} = \sqrt{\frac{(15-1)14.7^2 + (10-1)3.2^2}{15+10-2}} = 11.6$$

If we had not computed S.D.'s, or if the research report we were perusing presented the S.E.M. rather than the S.D., we can employ Eq. (5) to transform S.E.M. to S.D.,

$$\text{S.E.M. to S.D.} = (4.6)(\sqrt{10}) = 14.5 \text{ [Alzheimer sample]}$$

If we computed g and later decided we wanted d, we would employ Eq. (6)

$$g \text{ to } d = \sqrt[-5.78]{\frac{18}{20}} = -2.18$$

Conversely, if we computed d and later wanted g, we would employ Eq. (7)

$$d \text{ to } g = \sqrt[-2.07]{\frac{20}{18}} = -5.48$$

It should be obvious that the transformation of g to d and g to d is crude at best. It is always best to compute effect sizes based on raw data, and this notion keeps true for all effect size conversion formulae. Although the conversion formula will not alter the direction of the effect size erroneously, it will, and always will be, a clumsy estimate. Similarly, we can transform Hedges' g to r using Eq. (8)

$$r = \frac{-5.78}{\sqrt{-5.78^2 + 4\left[\frac{18}{20}\right]}} = -0.99$$

or Cohen's d to r using Eq. (9)

$$r = \frac{-2.07}{\sqrt{-2.07^2 + 4}} = -0.72$$

We can see from the above expressions using the more familiar correlative interpretation that Hedges' g corresponds to a very large effect indeed. Moreover, we can also see how Cohen's effect is more conservative as it does not assume equality of variance.

If the Boston Naming Test example was a published study that included no raw data, nor a description of the sample means or S.D.'s, but did include the inferential statistics t and F, we could employ the formulae found in Table 3 to obtain Cohen's d

$$t \text{ to } d = \frac{2(-3.89)}{\sqrt{18}} = -1.83$$

$$F \text{ to } d = \frac{2(15.16)}{\sqrt{18}} = -1.83$$

Although the conversion of t or F to d is reliable, again, we can see that it is only a crude estimate of the actual effect size in keeping with its computation from means and S.D.'s. Moreover, if we were interested in the transformation of our computed correlation between duration of illness and Boston Naming Test raw scores in patients with AD (see Table 4), we can also employ the r to d conversion formula presented in Table 3

$$r \text{ to } d = \frac{2(-0.96)}{\sqrt{1 - (-0.96^2)}} = -6.85$$

A large effect indeed.

If we return to Cohen's effect size obtained from Eqs. (2) and (3) (i.e., means and S.D.'s) for the hypothetical data set, we note that the overlap % is approximately 18. This OL% would indicate that approximately 80% of the patients with AD obtained scores unlike the normal controls. If we glance at Table 4, we can see that this is indeed the case. That is, only two patients with AD (20% of the sample) obtained scores that fell within the distribution of scores obtained by the normal controls (scores 55, 52). Conversely, 80% of the normal controls obtained scores that did not fall within the distribution of scores obtained by the patients with DAT.

Now, more importantly, we need to decide whether our obtained effect size is small, medium, or large. Cohen (1988) recognized that by supplying a common conventional frame of reference in the interpretation of the magnitude of *d*, where an effect of 0.2 corresponds to small, 0.5 medium, and 0.8 large, that he was taking a risk in offering conventional operational definitions for these terms for use in effect size analysis in as diverse a field of inquiry as behavioral science. Over the years, however, this risk has nevertheless been accepted in the belief that more is to be gained than lost by supplying a common conventional frame of reference for the interpretation of effect sizes. What seems important for neuropsychological researchers to note however, is that Cohen's heuristic benchmarks were formulated from his review of the treatment efficacy literature in psychology and his impressions of result typicality across a broad range of the social science literature. For example, the effects of treatment in phenylpyruvic mental deficiency. Note that ensuing meta-analysis findings suggest that his estimates were probably in the ballpark, but that typicality is a fact statement, not a value judgment of any kind. Furthermore, this fact statement may not generalize to other disciplines or to some subdisciplines. As neuropsychologists, we are indeed engaged in research of treatment efficacy in terms of its effect on cognition and behavior. We are also engaged in understanding the relationship between brain and behavior, however. Hence, is it appropriate to adopt a benchmark heuristically devised to aid only in the interpretation of social science broadly defined?

Let us say we were interested in the difference between patients with AD and normal controls on a measure of verbal learning. We posit that because AD is primarily a disease affecting medial temporal and parietal cortices, verbal learning should be impaired in patients with the disease. We would expect a verbal learning deficit to be present in keeping with the observation that lesions to this area typically result in such deficits. Hence, we administer a measure of verbal learning to a sample of patients with probable AD and normal controls. We compare the mean performance between groups and calculate an effect size of 1.0. This effect corresponds to approximately 45% overlap. Thus, about half of the patients with AD obtained scores that were unlike any of those obtained by the normal controls. According to Cohen's heuristic benchmark, this is a "large" effect. Now, let us suppose we wanted to test the efficacy of a particular drug for its purported prophylactic benefit in preventing cognitive decline in patients with chronic progressive multiple sclerosis (MS). We gather a sample of patients with MS and randomly assign patients to either the experimental group or the placebo group and take a baseline of cognitive status. We administer the drug to the experimental group and the placebo to the control group and after about a year, we do a second assessment of cognitive status. We compute an effect size between experimental and control sample means at follow-up and obtain a value of 0.25

corresponding to 80% overlap. Thus, 20% of the experimental group did not decline further in terms of cognition over the year. According to Cohen's heuristic benchmark, this is a "small" effect.

Now let us suppose we wanted to explore further our results in the previous hypothetical studies. In the first example, we determine that patients with AD are seemingly worse in terms of delayed recall relative to the normal controls. We therefore compute an effect size and obtain a value of 3.3 corresponding to less than 5% overlap. In our second experiment, we want to determine the "physical side-effect" profile of the drug, hence, we compare the experimental and control samples on the Expanded Disability Status Scale (EDSS; Kurtzke, 1983). We find an effect size of 0.01 (99% overlap).

With these hypothetical effects in mind, consider the conventional frame of reference for the interpretation of effect sizes provided by Cohen. The effect obtained ($d = 1.0$) for verbal learning was "large." Relative to the effect we obtained for delayed recall ($d = 3.3$), can it still be considered "large" as it can discriminate approximately all patients with AD from normal controls whereas verbal learning can discriminate about half? Moreover, the effect obtained ($d = 0.25$) for the efficacy of the drug in MS was "small" according to Cohen benchmarks. If the drug had prevented further decline of cognition in 20% of the patients, would patients with MS consider this effect "small," especially in the context of an absent side-effect profile where 1% of drug users will experience a great deal of added physical disability?

The crux I am trying to make is that the interpretation of an effect size should depend on research context rather than an inappropriate frame of reference. Although Cohen's benchmarks serve well in the evaluation of treatment effects in the field of general psychology, they are not always appropriate for neuropsychological research. What then is our alternative if Cohen's benchmarks are of limited appropriateness?

Given that we are accustomed to statistical algorithms that decide the "significance" of our results (e.g., $P < .05$), perhaps we do need some rule of thumb in the interpretation of effects. As such, I have provided such heuristic interpretation of effects in Table 1 that might seem particularly appropriate to neuropsychological research. That is, we can all agree that an effect size of 0.0 has no effect whatsoever. We can also surely agree that an effect size capable of discriminating approximately all experimental from control group participants in a two-group design (i.e., $d > 3.0$, OL% < 5) a "large" effect. For example, Zakzanis, Leach, and Kaplan (1999) have shown that several cognitive tests correspond to effect sizes greater than 3.0 across varying syndromes of dementia. Moreover, by way of double dissociation of effect sizes, we argued that these effects whose sensitivity approached 100%, can act as cognitive markers to aid in the differential diagnosis of dementia and neuropsychiatric syndromes. Hence, we might adopt $d > 3.0$, OL% < 5 as a heuristic marker in neuropsychological research for what it is worth and no more than that. Moreover, it is between the range of effects of 0.1 and about 2.0 where context should govern the interpretation of the effect size. The qualification of "small" and "medium" effects should be added only when the field of inquiry is understood in terms of effect sizes. That is, unless we begin to give due consideration to effect sizes in our neuropsychological studies, it will be difficult to determine how large one's obtained effects are within a specific domain of neuropsychological research without a frame of reference. An immediate and obvious resolution would be to carry out a meta-analysis to determine the average effect

size across existing studies. This would allow the neuropsychological researcher to place their findings against what is known for a particular neuropsychological condition in terms that are statistically comparable.

Once we have decided whether are obtained effect size is small, medium, or large, we can always employ Orwin's (1983) fail-safe *N* formula to determine just how robust our finding is. For example, if we were to conduct a meta-analysis of Boston Naming Test performance in patients with AD, and we found a total of 22 studies, yielding a mean effect size of 1.2, we could employ Eq. (11)

$$N_{\text{fs}} = \frac{22(1.2 - 0.2)}{0.2} = 110$$

This tells us that 110 studies supporting the null hypothesis of $d = 0.2$ (an insignificant effect in most contexts) is needed to overturn our obtained effect size across 22 studies of 1.2. Hence, in keeping with our hypothetical data set, we can see that our obtained effect is larger than that found in most studies. We could then speculate why. Duration of illness? Onset age? Moreover, if we had stopped at illustrating the statistical significance of our finding only without due consideration given to the magnitude of effect, we would have no footing to state that our obtained effect size is larger than that found in the literature, and no basis to speculate why. Indeed, it has been all too common to ignore the magnitude of effect of one's findings in neuropsychological research. Pity. The observational truth regarding a great deal of our findings likely remains fully undiscovered in the published literature.

3. Conclusions

It is crucial to realize that finding "nonsignificance" is not the same as finding "no effect," as it is quite possible for a meaningful effect to be present although the statistical test lacks sufficient power to detect it at the desired significance level due to a modest sample size or an imprecise research design. It has been argued elsewhere that due consideration be given to the magnitude of effect along with traditional tests of statistical significance (e.g., Schmidt, 1996; Zakzanis, 1998). Although this opinion has been voiced for almost 40 years, researchers have seldom reported effect sizes in their works. One plausible explanation for this neglect may be that few researchers have a clear idea of when or how to calculate and interpret them. Accordingly, I have reviewed the standard calculations of commonly employed effect sizes in two group designs and have shown how to adjust some familiar, and perhaps not so familiar, formulae. I have also put forth an argument to adopt Cohen's measure as an expression of effect size based on its apropos to neuropsychological research. It was also argued that the interpretation of the magnitude of the effect size should depend on context, and not on pre-established heuristic benchmarks. It was noted, however, that effect sizes greater than 3.0 (OL% < 5) might seem particularly appropriate when searching for markers in neuropsychological disorders.

In summary, if, as neuropsychologists, we think of the relationship between brain and behavior as the same as that between truth and reality, we must be equipped with statistical procedures that are coherent in terms of what we measure and what it represents. I believe that

this statistical procedure is effect size analysis, and without it, I believe we fail to tell the truth, the whole truth, and nothing but the truth.

Acknowledgments

Appreciation is expressed to Dr. B. Thompson for his insightful recommendations, revisions, and meticulous review of the manuscript. Donald S. Birtch is also to be acknowledged for his helpful assistance with the Equation 3.0 software.

References

American Psychiatric Association. (1994). *DSM-IV: diagnostic and statistical manual of mental disorders*. Washington, DC: American Psychiatric Association Press.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*, 1–29.

Bieliauskas, L. A., Fatenau, P. S., Lacey, M. A., & Roper, B. L. (1997). Use of the odds-ratio to translate neuropsychological test scores into real-world outcomes: from statistical significance to clinical significance. *Journal of Clinical and Experimental Neuropsychology*, *19*, 889–896.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Cohen, J. (1994). The earth is round ($P < .05$). *American Psychologist*, *49*, 997–1003.

Cooper, H. (1979). Statistically combining independent studies: a meta-analysis of sex differences in conformity research. *Journal of Personality and Social Psychology*, *37*, 131–146.

Cooper, H., & Hedges, L. V. (1994). *The handbook of research synthesis*. New York: Russel Sage Foundation.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*, 379–390.

Guttman, L. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data analysis*, *1*, 3–10.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Kaplan, E., Goodglass, H., & Weintraub, S. (1983). *Boston naming test*. Philadelphia: Lea and Febiger.

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

Kraemer, H. C., & Andrews, G. (1982). A nonparametric technique for meta-analysis effect size calculation. *Psychological Bulletin*, *91*, 404–412.

Kurtzke, J. F. (1983). Rating neurological impairment in multiple sclerosis. An Expanded Disability Status Scale (EDSS). *Neurology*, *33*, 1444–1452.

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

Olejnik, S., Algina, J. (in press). Measures of effect size for comparative studies: applications, interpretations, and limitations. C*ontemporary Educational Psychology.*

Orwin, R. G. (1983). A fail-safe *N* for effect size. *Journal of Educational Statistics*, *8*, 157–159.

Rosenthal, R. (1979). The 'file drawer' problem and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage.

Rosenthal, R. (1995). Writing meta-analytic reviews. *Psychological Bulletin*, *118*, 183–192.

Rourke, B. P., Costa, L., Cicchetti, D. V., Adams, K. M., & Plasterk, K. J. (1991). *Methodological and biostatistical foundations of clinical neuropsychology.* Lisse, The Netherlands: Swets and Zeitlinger.

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

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers. *Psychological Methods*, *1*, 115–129.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *Journal of Experimental Education*, *61*, 334–349.

Soper, H. V., Cichetti, D. V., Satz, P., Light, R., & Orsini, D. L. (1988). Null hypothesis disrespect in neuropsychology: dangers of alpha and beta errors. *Journal of Clinical and Experimental Neuropsychology*, *10, 255–270.*

Thompson, B. (1998). Review of what if there were no significance tests? In: L. Harlow, S. Mulaik, & J. Steifer (Eds.), *Educational and psychological measurement, 58,* 332–344.

Thompson, B. (1999a). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory and Psychology*, *9* (2), 165–181.

Thompson, B. (1999b). Journal editorial policies regarding statistical significance tests: heat is to fire as p is to importance. *Educational Psychology Review*, *11*, 157–169.

Thompson, B., & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent JCD research articles. *Journal of Counseling and Development*, *76*, 436–441.

Vacha-Hasse, T., Nilsson, J. E., Reetz, D. R., Lance, T. S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory and Psychology*, *10*, 413–425.

Valdois, S., Joanette, Y., Poissant, A., Ska, B., & Dehaut, F. (1990). Heterogeneity in the cognitive profile of normal elderly. *Journal of Clinical and Experimental Neuropsychology*, *12*, 587–596.

Wilkinson, L., & The APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: guidelines and explanations. *American Psychologist*, *54*, 594–604.

Wittgenstein, L. (1949). *Philosophical investigations*. Cambridge: Cambridge Univ. Press.

Wolf, F. M. (1986). *Meta-analysis: quantitative methods for research synthesis*. Newbury, London: Sage.

Zakzanis, K. K. (1998). Brain is related to behavior ($P < .05$). *Journal of Clinical and Experimental Neuropsychology*, *20*, 419–427.

Zakzanis, K. K., Leach, L., & Kaplan, E. (1999). *Neuropsychological differential diagnosis*. Lisse, The Netherlands: Swets and Zeitlinger.