

RUNNING HEAD: Primer on confidence intervals

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals based on central and noncentral distributions. *Educational and Psychological Measurement*, 61, 530-572.

© Sage Publications

Journal website: <http://www.sagepub.com/journal.aspx?pid=165>

This article may not exactly replicate the final version published in the journal. It is not the copy of record."

A primer on the understanding, use and calculation of confidence intervals that are based  
on central and noncentral distributions

Geoff Cumming  
La Trobe University  
Sue Finch  
The University of Melbourne

Address correspondence to:

Geoff Cumming

School of Psychological Science, La Trobe University

Bundoora, Australia 3086

Email: G.Cumming@latrobe.edu.au

---

*Author note.* The authors thank Fiona Fidler and Michael Smithson for valuable comments on drafts of this paper, and Geoff Robinson for the routines used to calculate noncentral  $t$  in the *ESCI* software referred to in the paper. The authors may be contacted by email: G.Cumming@latrobe.edu.au. Information about the software, which runs under Microsoft Excel, and its availability may be obtained at [www.psy.latrobe.edu.au/esci](http://www.psy.latrobe.edu.au/esci).

### Abstract

Reform of statistical practice in the social and behavioural sciences requires wider use of confidence intervals (CIs), effect size measures and meta-analysis. We discuss four reasons for promoting use of CIs: they (i) are readily interpretable; (ii) are linked to familiar statistical significance tests; (iii) can encourage meta-analytic thinking; and (iv) give information about precision. We discuss calculation of CIs for a basic standardised effect size measure, Cohen's  $\delta$  (also known as Cohen's  $d$ ), and contrast these with the familiar CIs for original score means. CIs for  $\delta$  require use of noncentral  $t$  distributions, which we apply also to statistical power and simple meta-analysis of standardised effect sizes. We provide the *ESCI* graphical software, which runs under Microsoft Excel, to illustrate the discussion. Wider use of CIs for  $\delta$  and other effect size measures should help promote highly desirable reform of statistical practice in the social sciences.

The popular media make statements like these:

Support for the Prime Minister was 38% in a poll with an error margin of 4%.

The last ice age ended between 11,500 and 10,000 years ago.

Such statements are, at an informal level, readily comprehensible. They provide not only a ‘most likely’ value of the variable of interest, but also information about our uncertainty or the precision of estimation.

More formally, many scientific disciplines routinely report findings by making statements such as:

The speed of conduction of the diseased nerves was  $1.25 \pm 0.4 \text{ ms}^{-1}$ .

A kilometre downwind the concentration of the pollutant was 14 ppm, with 95% confidence interval (7, 27).

We are given a point estimate and an interval estimate, and once again the information is—at least in an informal, intuitive way—easy to understand.

Each of these statements refers to a confidence interval (CI), which is an interval or range of plausible values for some quantity or population parameter of interest. A CI is a set of parameter values that are reasonably consistent with the sample data we have observed. As the examples illustrate, CIs provide a mechanism for making statistical inferences that give information in units with practical meaning for both the researcher and the reader. They give a best point estimate of the population parameter of interest, and an interval about that to reflect likely error—the precision of the estimate.

Although our examples do not all make it explicit, the description of the interval is accompanied by a statement of the confidence level, usually expressed as a percentage, such as ‘95% CI’. We refer to this percentage as  $C$ . It is never certain (unless we use  $C$

= 100) that the interval includes the true value of the parameter of interest. The confidence level, or probability, describes the chance of intervals of this kind including, or ‘capturing’, the population value in the long run. We discuss below the correct interpretation of this probability.

Imprecision as represented by the width of the CI can come from a variety of sources, including sampling error (the number of people asked about their support for the Prime Minister was not very large) and error of measurement (speed of nerve conduction was measured by an indirect, inaccurate technique). The width of the interval also depends on  $C$ , the probability level chosen—other things being equal, higher probability levels result in wider intervals.

#### **Four reasons to use confidence intervals**

Our aim in this paper is to discuss four main reasons for using CIs, to present formulas and examples, and to go beyond intervals for means based on the familiar central  $t$  distribution to intervals for standardised effect sizes based on the noncentral  $t$  distribution. We provide the *ESCI* software, which runs under Microsoft Excel, to support the discussion.

Four main reasons for using CIs are:

1. They give point and interval information that is accessible and comprehensible and so, as the examples above illustrate, they support substantive understanding and interpretation.
2. There is a direct link between CIs and familiar null hypothesis significance testing (NHST): Noting that an interval excludes a value is equivalent to rejecting a

hypothesis that asserts that value as true—at a significance level related to  $C$ . A CI may be regarded as the set of hypothetical population values consistent, in this sense, with the data.

3. CIs are useful in the cumulation of evidence over experiments: They support meta-analysis, and meta-analytic thinking focussed on estimation. This feature of CIs has been little explored or exploited in the social sciences but is in our view crucial, and deserving of much thought and development.
4. CIs give information about precision. They can be estimated before conducting an experiment and the width used to guide the choice of design and sample size. After the experiment they give information about precision that may be more useful and accessible than a statistical power value.

### **Statistical reform and advocacy of confidence intervals**

The American Psychological Association's (APA) Task Force on Statistical Inference (TFSI) recently advocated the increased use of CIs in reporting the results of psychological studies (Wilkinson & TFSI, 1999). For example, CIs should be reported along with, or instead of, hypothesis test results: "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" (p. 599).

The Task Force emphasised the importance of presenting findings in metrics that can be directly interpreted, and of calculating CIs in these metrics:

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)

This statement relates to our Reason 1. As noted, an important aspect of choice of metric is the choice between original and standardised units for reporting an effect size and its CI; we take up that question in detail below.

The Task Force continued with remarks related to our Reasons 3 and 4:

We must stress again that reporting and interpreting effect sizes in the context of previously reported effects is essential to good research. It enables readers to evaluate the stability of results across samples, designs, and analyses. Reporting effect sizes also informs power analyses and meta-analyses needed in future research. ... Interval estimates should be given for any effect sizes involving principal outcomes. ... Comparing confidence intervals from a current study to intervals from previous, related studies helps focus attention on stability across studies ... (p. 599)

These recommendations promote an important move away from a long-standing overreliance on NHST as *the* method for making statistical inferences in psychology. Finch, Thomason and Cumming (2001) discussed further aspects of the debate and efforts towards reform, in particular the role of past and future editions of the APA *Publication Manual* (e.g., APA, 1994). As Kirk (2001) noted in his recent EPM article:

The APA publication manual and similar manuals are the ultimate change agents. If the 1994 edition of the APA manual can tell authors what to capitalize, how to reduce bias in language, when to use a semicolon, how to abbreviate states and

territories, and principles for arranging entries in a reference list, surely the next edition can provide detailed guidance about good statistical practices. (p. 217)

And the Chair of the APA Publications Committee coordinating the on-going revision of the Publication Manual observed that

Given the instructions on statistical reporting in the current [4th edition] Publication Manual, it is scarcely a radical change to provide instructions to report effect sizes... [I]t is unfortunate that the reporting of effect sizes has been framed as a controversy. Reporting of effect sizes is, instead, simply good scientific practice. (Hyde, 2001, pp. 227-228)

Despite a long history of criticism and exhortation to adopt a wider variety of alternative exploratory, descriptive and inferential methods, NHST remains the dominant technique (e.g., Bakan, 1966; Cohen, 1994; Loftus, 1993; Meehl, 1978; Nickerson, 2000). We strongly support these calls for reform, and believe that wider understanding and use of CIs should be a central aspect of changes to statistical practice in psychology, education and cognate disciplines.

#### *Wider use of confidence intervals*

CIs were introduced by Neyman in the 1930s (Cowles, 1989) and have been advocated in the psychological literature at least since the late 1950s (e.g., Chandler, 1957; Grant, 1962; LaForge, 1967). Before formal inferential methods were widely used in psychology the probable error was often reported. The probable error “is that deviation from the mean of a normal distribution that corresponds to a point dividing the area between the mean and tail into two equal halves” (Gigerenzer & Murray, 1987, p.18), that is, approximately two-thirds of a standard deviation. A criterion of 3 times the

probable error was often used. Although this corresponded approximately to a 95% CI, it was often interpreted as an indication that results 'deviated from chance'.

Many psychologists and educators are likely to have once learnt about simple procedures for calculating CIs for means, however CIs are not typically reported in psychological journals (Finch, Cumming & Thomason, 2001). A number of explanations have been suggested for psychologists' failure to calculate and report CIs. Schmidt and Hunter (1997) presented arguments to counter a number of reasons given for continued use of NHST, and speculated on why there is strong resistance to change:

Accepting the proposition that significance testing should be discontinued and replaced by point estimates and confidence intervals entails the difficult effort of changing the beliefs and practices of a lifetime. Naturally such a prospect provokes resistance. Researchers would like to believe there is a legitimate rationale for refusing to make such a change. (p. 49)

The tradition of NHST in psychology may also have meant that psychologists remained largely ignorant of CIs, especially as CIs have not been readily available in software commonly used in the social sciences (Steiger & Fouladi, 1997). Further, "interval estimates are sometimes embarrassing" (Steiger & Fouladi, p. 228)—that is, they reveal that many psychological studies are very imprecise.

Reformers have argued that psychologists should routinely report CIs, which can be done in a variety of ways. In medicine it is common practice to report numerical values of CIs in tables or text; in some sciences it is standard to report means in a figure, and attach error bars, which are commonly either the standard error of the means or 95% CIs. (Standard error bars can be regarded as giving an approximate 68% CI if the sample

size is not small.) In some fields of psychology, such as psychobiology, one or more of these practices are fairly common, but across much of published psychological research there is very little or no use of CIs, and much room for improvement of statistical practices.

We have some evidence to suggest that the wider uptake of CIs in psychology will require more than just the adding of the Task Force recommendations to a revised APA Publication Manual. As editor of Memory and Cognition between 1994 and 1997, Geoffrey Loftus asked authors to report their results as figures with error bars rather than relying on statistical significance tests. As we report elsewhere (Finch, Cumming, Williams, et al., 2001), authors found it difficult to relinquish their reliance on statistical significance testing and were often unable or uncertain about how to calculate appropriate error bars (or CIs) for their particular research designs. Also, even when they did follow Loftus' request and include a figure with error bars, only rarely were these bars used to assist interpretation. Loftus' further request, to use error bars to obviate the need for NHST and *p*-values, was almost never followed. Finch et al. concluded that, while editorial requirements may be important in achieving improved statistical practice, other measures are also needed. They noted, however, the greater success of reforms in medicine where many editors adopted standard recommendations for reporting statistical information, including use of CIs.

A further recommendation of the Task Force was that graphical representation should be used freely:

Although tables are commonly used to show exact values, well-drawn figures need not sacrifice precision. Figures attract the reader's eye and help convey

global results. Because individuals have different preferences for processing complex information, it often helps to provide both tables and figures. This works best when figures are kept small enough to allow space for both formats. Avoid complex figures when simpler ones will do. In all figures, include graphical representations of interval estimates whenever possible. (Wilkinson & TFSL, 1999, p. 601)

We concur, and go further: We believe that vivid graphical representations can assist greatly in the understanding of statistical concepts. It can be even better if these representations are dynamic and interactive (Thomason, Cumming & Zangari, 1994). Therefore in designing the *ESCI* software we aimed for maximum use of graphics and interactivity.

### **Aims of this paper**

Our aims here are to provide explanations and resources that will assist social and behavioural scientists in understanding and using CIs for some simple research designs. We focus on the following basic cases:

1. Single group design, where interest is in the population mean;
2. Two independent groups design, where interest is in the difference between the two population means; and
3. Single group, two repeated measures (e.g., pretest and posttest measures) design, where interest is in the population mean for change from first to second measurements.

Recent papers in the psychological literature have described procedures for calculating CIs for effect size measures, and have explained that in many situations *noncentral* distributions are needed for calculating CIs (e.g., Fidler & Thompson, 2001; Smithson, 2001; Steiger & Fouladi, 1997). We aim in this paper to provide an elementary introduction as a basis for the more complex cases taken up in those papers, which include a range of multiple regression and analysis of variance models.

We will describe when central or noncentral distributions are required for calculation of CIs for selected population parameters of interest to social scientists. We will refer to the *ESCI* tools that allow users to investigate some concepts related to CIs, and to use CIs with simple data sets.

## CASE 1: SINGLE GROUP DESIGN

### **The basic confidence interval**

The  $t$ ,  $F$  and  $\chi^2$  distributions are *families* of distributions. For example, there is not a single  $t$  distribution, but rather many  $t$  distributions that vary in spread depending on the degrees of freedom. Although the most common application of these distributions in psychology and education has been for NHST, they also underpin the calculation of CIs. Here we will remind readers of the application of the familiar  $t$  distribution for calculation of CIs for a mean in original measurement units.

Our example is for Case 1, the simple case of calculating the CI for the mean of a single group, of size  $n$ , assumed to be a random sample from a normally distributed population. Consider:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \quad [1]$$

where “ $\sim$ ” means ‘is distributed as’. This quantity can be thought of as a standardisation of the difference between an estimator, the sample mean  $\bar{X}$ , and the parameter, the population mean  $\mu$ . Irrespective of the value of  $\mu$ , the quantity above has a  $t$  distribution with  $(n - 1)$  degrees of freedom. It has a symmetric  $t$  distribution centred about zero. This symmetry arises because  $(\bar{X} - \mu)$  is normally—and therefore symmetrically—distributed about zero. The sample standard deviation  $S$  (in the denominator) is closely related to the  $\chi^2$  distribution.

#### *Confidence intervals, Method 1*

First we describe the conventional explanation of the basic CI formula. We will refer to this approach as Method 1.

We can make probability statements about the quantity in [1], such as the following:

$$\Pr(-t_{n-1}(0.975) < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{n-1}(0.975)) = 0.95 \quad [2]$$

where  $t_{n-1}(0.975)$  refers to the 0.975 quantile from a  $t$  distribution with  $(n-1)$  degrees of freedom. It is the value of the  $t$  distribution that gives upper and lower tails that sum to 5% of the area under the distribution. Note that in situations where  $n$  is large, the 0.975 quantile from the  $t$  distribution is very close to that of the standard normal distribution, a

value of 1.96. So for 95% of large samples, of any given size, the distance between any sampled mean and  $\mu$  will be no more than 1.96 times the standard error. Hence [2] takes a familiar form for large samples:

$$\Pr(-1.96 < \frac{\bar{X} - \mu}{S/\sqrt{n}} < 1.96) \approx 0.95 \quad [3]$$

and the approximation becomes closer as  $n$  increases.

Rearrangement of [2] provides the familiar result for a CI. First, we multiply through by  $S/\sqrt{n}$ :

$$\Pr(-t_{n-1}(0.975) \times S/\sqrt{n} < \bar{X} - \mu < t_{n-1}(0.975) \times S/\sqrt{n}) = 0.95 \quad [4]$$

This probability statement refers to the distance between a sample mean and the population mean that it is intended to estimate, measured in terms of the standard error of the mean:  $S/\sqrt{n}$ .

Further rearrangement gives a probability statement with an interval centred around  $\mu$ :

$$\Pr(\bar{X} - t_{n-1}(0.975) \times S/\sqrt{n} < \mu < \bar{X} + t_{n-1}(0.975) \times S/\sqrt{n}) = 0.95 \quad [5]$$

This probability statement gives directly the following formula that we can apply to our data to calculate a 95% CI for  $\mu$ :

$$\bar{x} \pm t_{n-1}(0.975) \times s/\sqrt{n} . \quad [6]$$

(We are following the convention that an uppercase Roman letter signals a random variable, and the corresponding lowercase letter is a particular realisation of that variable. So in the earlier general probability statements we used  $\bar{X}$  and  $S$ , but here we use  $\bar{x}$  and  $s$  because these are the values calculated from a sample.) Essentially, Method 1 for

constructing a CI for  $\mu$  involves rearranging a probability statement about  $\bar{X}$  to make a probability statement about an interval containing  $\mu$ .

*Pivotal quantities*

Quantities of the general form

$$\frac{(\text{Estimator} - \text{Parameter})}{SE}$$

are pivotal quantities. Probability statements using pivotal quantities can typically be ‘pivoted’ or rearranged to construct CIs for population parameters. The Case 1 quantity in [1] is pivotal, and we illustrated above how it can be pivoted to obtain the CI for  $\mu$ . Similar principles can be used to derive CIs for population mean differences and population variances.

We emphasise that the procedure for calculating the CI for  $\mu$  does not depend in any way on the particular value of the population parameter  $\mu$  that we are estimating. This is a property of pivotal quantities: Estimation of the bounds of the CI does not depend on the value of the parameter we wish to estimate (Cox & Hinkley, 1974). We will see below that CIs for effect sizes are typically more difficult to estimate, because such intervals depend on quantities that are not pivotal.

**An original units example, Case 1**

*Example* A test of verbal ability has been constructed to have a population mean score of 25. We take a sample of 12 children and assess their verbal ability scores as {33.0, 47.0, 23.5, 35.0, 35.5, 26.0, 28.5, 24.0, 37.0, 34.0, 30.0, 38.0}. We calculate  $\bar{x} = 32.6$  and  $s = 6.72$ .

Figure 1 shows a simple dot plot of the data points for this single group, the sample mean and the 95% CI for the population mean  $\mu$ , which is (28.4, 36.9). (You can use *ESCI* **CIoriginal**<sup>1</sup> to make these calculations and generate a similar display for your own data.) Variations on Figure 1 (or explorations with **CIoriginal**) illustrate some basic features of the original units CI for the population mean, including:

- The CI is centred on the sample mean and is symmetric.
- Higher chosen  $C$  (percent confidence) requires a wider CI.
- Larger  $n$  (sample size) gives a shorter CI.
- For many small but realistic datasets in psychology, the 95% CI often may seem disappointingly wide.

### **Confidence intervals as a set of plausible values for $\mu$**

Figure 1 also shows  $\mu_0$ , which is any particular value we may care to consider (25, in the case illustrated) for the population mean. If we regard this value as being specified by a null hypothesis, we can conduct a conventional  $t$  test of our sample mean against this value; the  $t$  and two-tail  $p$ -value for that test are displayed by **CIoriginal**. Varying  $\mu_0$  illustrates the relationship between a CI and the  $p$ -value: The interval captures  $\mu_0$  if and only if the  $p$ -value, expressed as a percentage, is greater than  $(100 - C)$ , where  $C$  is our chosen percent confidence level. We can take the  $p$ -value as an index of how far our sample mean is from any particular  $\mu_0$  and can regard our sample mean as ‘compatible with’ that  $\mu_0$  if the  $p$ -value is not too small. Further, the CI is just that set of  $\mu_0$  values meeting this criterion. Informally, the CI is the set of  $\mu$  values that could plausibly have given our sample mean.

### *Confidence intervals and NHST*

The informal argument of the previous paragraph can, and usually is, expressed in NHST terms: The  $C\%$  CI captures  $\mu_0$  if and only if the conventional test of  $H_0: \mu = \mu_0$  (at the  $(100 - C)$  level) is not statistically significant, so the CI is that set of  $\mu$  values for which the data would not lead to rejection of the corresponding null hypotheses. Although our Reason 2 is the link between CIs and NHST, in the presentation of Method 1 above we avoided NHST language to show that understanding of CIs need *not* depend on NHST, and as part of the general reform agenda to de-emphasise NHST. In addition, note that a CI can be computed even (a) if no null hypothesis is stated, or (b) if a stated null hypothesis turns out to posit a wildly wrong parameter value.

Reason 2, the link between CIs and NHST, should not be used simply to tie every use of CI back to a statistical significance test. As Thompson (1998) argued, "If we mindlessly interpret a confidence interval with reference to whether the interval subsumes zero, we are doing little more than nil hypothesis statistical testing" (p. 800). We have observed such examples in practice (Finch, Cumming, & Thomason, 2001). There are other aspects to this relationship. For example, the best way to teach NHST may be to teach CIs thoroughly first, then to introduce, explain and motivate NHST in terms of the relationship with CIs. Hunter (1997) advocated this approach, which is used in some popular statistics textbooks (e.g., Moore & McCabe, 1993). It is also commonly used in statistics textbooks in the disciplines of, for example, management and business. Lockhart (1998) made much use of CIs in his statistics text for psychology and, most recently, Smithson's (2000) text *Statistics with confidence* is based strongly on CI ideas.

*Confidence intervals, Method 2*

Thinking of the CI as the set of plausible population parameter values leads to Method 2, an alternative derivation of the CI formulas given above. This can serve as a bridge to the derivation of CIs using noncentral distributions we consider below.

Consider  $\mu_L$  and  $\mu_U$  to be the two most extreme values of  $\mu$  that would meet our criterion of plausibility, or compatibility with our data, as shown in Figure 2. These are of course simply the ends of the CI, so

$$\mu_L = \bar{x} - t_{n-1} \times s / \sqrt{n} \quad \text{and} \quad \mu_U = \bar{x} + t_{n-1} \times s / \sqrt{n}. \quad [7]$$

Our criterion means that  $\mu_L$  is chosen so that the right tail area of the sampling distribution of  $t$  that lies to the right of  $\frac{\bar{x} - \mu_L}{s / \sqrt{n}}$  is  $\frac{1}{2}(100 - C)/100$ , and correspondingly for  $\mu_U$ .

In NHST terms, and referring as in Figure 2 to the sampling distributions scaled to the original units axis,  $\mu_L$  and  $\mu_U$  are such that our sample mean lies on the boundary of the rejection region, for null hypotheses of  $\mu_0 = \mu_L$  and  $\mu_0 = \mu_U$ . The critical distance, which gives the length of each arm of the CI, is that from  $\mu_L$  to our mean (or  $\mu_U$  to our mean). Because in this case, as noted earlier, the shape of the sampling distribution is independent of  $\mu$  (that is,  $t$  for the original units sample mean is pivotal), this distance is the same as that for the sampling distribution around any  $\mu$ .

To put it another way, and using  $C=95$ , we wish to find the lower and upper bounds on  $\mu$  such that:

$$\Pr(t_{n-1} \geq \frac{\bar{x} - \mu_L}{s / \sqrt{n}}) = 0.025 \quad \text{and} \quad \Pr(t_{n-1} \leq \frac{\bar{x} - \mu_U}{s / \sqrt{n}}) = 0.025. \quad [8]$$

Cox and Hinkley (1974) described what they call a simple rule: “to obtain ‘good’  $1 - \alpha$  upper confidence limits, take all those parameter values not ‘rejected’ at level  $\alpha$  in a ‘good’ significance test against lower alternatives. There is an obvious modification for obtaining ‘good’ lower confidence limits” (p. 214).

To summarise, Method 2 is the following process for obtaining the CI: Thinking graphically and with Figure 2 in mind, slide the sampling distribution until the right tail beyond  $\bar{x}$  is the criterion size of  $\frac{1}{2}(100 - C)/100$ ; the mean of the distribution then gives the lower bound of the interval. Correspondingly find the upper bound. The calculation is easy because the relevant distance is independent of  $\mu$ , the parameter being estimated. Below we will need to use Method 2 to derive CIs for a standardised effect size measure, and in that case the relevant distance does depend on the parameter being estimated.

### **The level of confidence, $C$**

Figure 3, which comes from **CIjumping**, may assist understanding of ‘level of confidence’. It shows an assumed normal population, mean  $\mu$ , the dot plot of a random sample, size  $n$ , and the mean and 90% CI for  $\mu$  calculated from that sample. It also shows the means and CIs calculated from 19 previous samples taken independently from the population. In any real case we have just one sample mean, and thus one CI. But this can be considered one of a potentially infinite set of intervals, each generated from an independent replication of the experiment. In the long run,  $C\%$  of intervals would cover  $\mu$ , the population mean.  $C$ , considered as a probability, applies to the process of calculating the CI, rather than to any particular interval. In thinking about the meaning of

level of confidence, note that it is the CI that varies, while  $\mu$  is fixed although unknown, as illustrated in Figure 3.

Considering Figure 3, and **CIjumping**, it is important to recognise that the situation being represented is artificial in two ways: In any real experimental situation (i) we do not know  $\mu$ , and (ii) we take only one sample, not many.

Even so, exploration of the simulation illustrates a number of important features of CIs. For example:

- In accord with the formula, average CI width varies inversely as the square root of  $n$ : To halve the average width we need to take samples four times as large.
- Observing sequences of CIs often gives the impression that successive independent CIs jump around to a surprisingly large extent, and very haphazardly. Realising this may help undermine the compelling illusion, related to the Law of Small Numbers (Tversky & Kahneman, 1971), that our particular sample and CI must surely represent the population closely!

### **The standardised effect size, Cohen's $\delta$**

We earlier quoted some statements by the TFSI about effect sizes. Reporting effect size measures is important because it focuses attention on the substantive results and facilitates meta-analysis. Often an original score effect size will be most readily understood, but a standardised effect size should be reported as well if doing so helps research communication. We will consider here Cohen's  $\delta$ , a simple and basic standardised effect size measure. (Note that Cohen (1988) referred to this measure as  $d$ ; we discuss our choice of notation in a section below.) Cohen (1988) and Hunter and

Schmidt (1990), among others, have explained how  $\delta$  is related to Pearson's correlation  $r$ , and to some other effect size measures. Our discussion just of  $\delta$  may therefore have some breadth of applicability.

Informally, Cohen's  $\delta$  is a mean or mean difference, standardised via division by a relevant standard deviation. We follow Hunter and Schmidt's (1990) notation and use  $\delta$  for the population parameter, and  $d$  when we calculate the sample statistic. For our Case 1, we define

$$\delta = \frac{\mu - \mu_0}{\sigma} \quad [9]$$

where  $\mu_0$  is a reference value for the population mean, often but not always chosen to be zero. The corresponding sample statistic that we calculate is

$$d = \frac{\bar{x} - \mu_0}{s} \quad [10]$$

In this single sample case, the conventional  $t$  test statistic for testing the hypothesis  $\mu = \mu_0$  about the population mean, with  $(n-1)$  degrees of freedom, is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad [11]$$

The last two formulas immediately give the relation that, for Case 1,

$$d = t/\sqrt{n} \quad [12]$$

This simple and appealing formula is worth remembering: Even if you are working with statistical software that does not give  $d$ , it is sure to give  $t$ , then for Case 1 simply dividing by  $\sqrt{n}$  gives  $d$ .

Figure 4, from **CIdelta**, shows as in Figure 1 our example data as an original score dotplot, and  $\bar{x}$  and the CI for  $\mu$ . It shows also the scale of  $d$ , which has zero at the

chosen  $\mu_0$  reference point of 25 and is in units of the standard deviation  $s$ , and the value of  $d$ , which is calculated as  $d = (\bar{x} - 25)/s = 1.14$ . Note that we are using 25, the known  $\mu$ , as the reference value for calculating the effect size, but are using  $s$  to estimate the unknown  $\sigma$ . The  $d$  value is a simple descriptive statistic for this group of children: Their mean verbal ability score is a little more than one standard deviation higher than 25. (We might choose that reference value because we are interested in assessing how different our group of children is, on average, from the mean of the population on which the test was developed.) By ‘effect size’ here we refer to the difference between the mean for this group and our chosen reference value of 25.

If we were prepared to regard our group as a random sample from some population of children, our  $d$  would be an estimate of  $\delta$  in that population, and the  $t$  statistic (shown by **C|original**; see Figure 1) could be used to test the hypothesis that the mean in this population equals 25.

Cohen (1988) emphasised that an effect size should be interpreted in its context, and argued that, because  $d$  is in standard deviation units, it can be interpreted as having general meaning. Cohen suggested that, in many situations in behavioural science,  $d$  values of 0.2, 0.5 and 0.8 can be regarded, respectively, as ‘small’, ‘medium’, and ‘large’. He cautioned, however, that these values are arbitrary and that effect sizes should be interpreted in their particular research situation. If we follow Cohen’s suggestion, the  $d$  in the example above is a very large effect. Cohen’s  $d$  has value as a descriptive statistic, and also as a measure that can be used in meta-analysis to combine the results from studies that used a variety of original measures.

We next wish to find the CI for  $\delta$ , but this turns out to be complex because  $d$  (like  $t$ , but *unlike* the simple original score mean  $\bar{X}$ ) is a ratio of two quantities, each of which is an estimate calculated from the data. This means that we must now consider the noncentral distributions.

### **Noncentral distributions**

The families of  $t$ ,  $F$  and  $\chi^2$  distributions are more extended than social scientists may realise. Typical inferential techniques are based on central distributions, like the  $t$  distribution described above. There are however *noncentral* distributions, distributions that are not centred at zero, that also prove useful for statistical inference. We next focus on noncentral  $t$  distributions and their application.

Central  $t$  distributions are described by one parameter, the degrees of freedom ( $df$ ). Non-central  $t$  distributions have an additional parameter, the noncentrality parameter, for which we use the symbol  $\Delta$ . (This parameter should not to be confused with the standardised effect size of Glass, 1976, which also uses this Greek letter.) Central  $t$  distributions, which are always symmetric, arise when a normally distributed variable with a mean of zero is divided by an independent variable closely related to the  $\chi^2$  distribution. (See the paragraph above containing [1].) Noncentral  $t$  distributions arise when a normally distributed variable with mean *not* equal to zero is divided by an independent variable closely related to the  $\chi^2$  distribution. They are not symmetric (see Figure 5) and the degree to which they are skewed depends on  $\Delta$ , the distance by which the mean of the normal distribution is displaced from zero.

## Notation

Before exploring the properties of noncentral  $t$  distributions, we should explain our choice of notation. A fundamentally important contemporary development in statistics in the social sciences is that of meta-analysis. Meta-analytic techniques are advancing rapidly and becoming more widely used. Terminology is becoming more consistent but has not yet fully stabilised. Even the symbols for describing Cohen's  $\delta$  are not yet universally agreed. Cohen (1988) originally defined  $d$  to be the population parameter, and used  $d_s$  for the sample statistic. Hunter and Schmidt (1990, and in their other influential writings on meta-analysis) used  $d$  for the sample statistic and  $\delta$  for the corresponding population parameter. This usage has the important advantage of being consistent with the widespread and valuable practice of distinguishing sample statistics (Roman letters:  $\bar{X}$ ,  $S$ ) from population parameters (Greek letters:  $\mu$ ,  $\sigma$ ). We adopt and recommend this latter usage rather than Cohen's.

However  $\delta$  is also used in mathematical statistics for the noncentrality parameter for  $t$  and other noncentral distributions. One argument of this paper is that standardised effect size measures, such as  $\delta$  should be used more widely, and in particular CIs for  $\delta$  should be reported where appropriate. This, as we shall see, requires use of noncentral  $t$ , and so we cannot avoid Cohen's  $\delta$  and noncentrality parameters being considered together. Thus we cannot use the symbol  $\delta$  in both roles. We elect to privilege meta-analytic considerations, use  $\delta$  as the population parameter for Cohen's effect size measure, and choose  $\Delta$  for the noncentrality parameter. In choosing  $\Delta$  we are following the authoritative if isolated precedent of Pearson and Hartley (1972). We make this choice despite the disadvantage that Glass (1976) used  $\Delta$  for his standardised effect size.

### Noncentral $t$ distributions

Figure 5 shows a number of noncentral  $t$  distributions and, for comparison, central  $t$  distributions. Until recently accurate calculation of noncentral  $t$  probabilities has been very difficult, and so particular attention has been paid by statisticians to numerical approximations and to the generation of tables for practical use. This intractability is shared by other noncentral distributions and is no doubt one reason why psychologists have made little use of them. Now, however, while development of accurate software remains challenging (Steiger & Fouladi, 1997), basic noncentral functions are provided in some statistical software widely used by social scientists, notably *SPSS*. In addition our *ESCI* software is intended to assist understanding of noncentral  $t$  and its use, and to provide a computation facility for simple cases.

Owen (1968) surveyed the properties and applications of noncentral  $t$ . Hogben, Pinkham and Wilk (1961) studied the mean and variance: The mean is  $a\Delta$ , where the factor  $a$  depends on  $df$  and approaches 1 as  $df$  increases. For  $df = 2$ ,  $a = 1.77$ ; for  $df = 15$ ,  $a = 1.05$ ; and for  $df = 60$ ,  $a = 1.013$ . Therefore noncentral  $t$  should appear centred approximately at  $\Delta$ , except for small  $df$ . The variance is a complex function involving  $\Delta^2$ . It approaches 1 as  $df$  increases, but only relatively slowly. It is a general property of noncentral  $t$  that it approaches its asymptotic symmetric—and in fact normal—shape only slowly.

**NonCentralt** supports exploration of how changes to  $df$  and  $\Delta$  influence the shape of the distribution. Figure 5 gives some examples to illustrate the effects of such changes.

Investigation with **NonCentral** illustrates properties of noncentral  $t$  distributions including:

- When  $\Delta = 0$ , noncentral  $t$  reduces to central  $t$ , and thus is symmetric and centred at zero.
- Noncentral  $t$  is centred approximately at  $\Delta$ , which may take any positive or negative value.
- For a given  $\Delta$ , as  $df$  increases, noncentral  $t$  approaches central  $t$  (and thus the normal distribution) in shape.
- The approach to the asymptotic symmetric shape is fairly slow: For  $\Delta = 2$ , even for  $df = 60$ , noncentral  $t$  is (just) visibly skewed (Figure 5). For larger  $\Delta$ , the approach to a symmetric shape is even slower. Therefore the rule of thumb often used by psychologists in other contexts that, when  $n$  is at least 30 the normal distribution is a sufficiently good approximation, should generally not be applied to noncentral  $t$ .
- The curve is skewed, very strongly for small  $df$ , and less so as  $df$  increases.
- The degree of skew increases markedly with the absolute value of  $\Delta$ .
- For positive  $\Delta$ , the skew is positive: The upper tail is fatter.
- Positive and negative  $\Delta$  values, of the same absolute value, give curves that are the same if reflected about a vertical through zero. For negative  $\Delta$ , the skew is negative. In other words, the outward tail (the tail remote from zero) is the fatter tail in every case.

### **Confidence intervals needing a noncentral $t$ distribution**

Consider the quantity

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1, \Delta}. \quad [13]$$

This quantity, the familiar one-sample  $t$  statistic, has in general a noncentral  $t$  distribution.

Only if  $\mu = \mu_0$  does it have a central  $t$  distribution. One way of thinking about the

location shift involved is to rewrite the statistic as:

$$\frac{\bar{X} - \mu + (\mu - \mu_0)}{S/\sqrt{n}} \sim t_{n-1, \Delta}. \quad [14]$$

The location shift  $(\mu - \mu_0)$  represents some difference between the true value of  $\mu$  and some chosen reference value  $\mu_0$ . We can regard this difference as a measure of the size of some effect.

So why is a central  $t$  distribution always used when conducting a null hypothesis test? Recall that in testing a null hypothesis it is always *assumed* that the null hypothesis is true. Hence, if the null value is  $\mu_0$ ,  $\mu = \mu_0$ , there is no shift in the distribution, and so the relevant distribution remains the central  $t$  distribution.

The noncentrality parameter for the distribution of  $t$  described above is:

$$\Delta = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}. \quad [15]$$

As we shall explain, it is estimation of this parameter that is necessary for the estimation of CIs for some standardised effect size measures.

The noncentrality parameter is closely related to the population effect size  $\delta$ :

$$\delta = \frac{\mu - \mu_0}{\sigma} = \Delta/\sqrt{n}, \quad \text{and so} \quad \Delta = \delta\sqrt{n}. \quad [16]$$

Hence we can find a CI for  $\delta$ , if we can find a CI for  $\Delta$ .

### Confidence intervals for Cohen's $\delta$

We first consider a CI for  $\Delta$ . Constructing a CI for  $\Delta$  is not straightforward because  $\Delta$  is a function of two parameters,  $\mu$  and  $\sigma$ , and both these must be estimated from the data. We cannot simply pivot a probability statement as we did for the CI for  $\mu$ , but must use the noncentral  $t$  distribution.

We can, however, think of the range of plausible values of  $\Delta$ , just as we considered plausible values of  $\mu$  in Method 2 for finding the CI for  $\mu$ . We define  $\Delta_L$  and  $\Delta_U$  to be the two extremes of the set of plausible values of  $\Delta$ , given our data; they will be the bounds of the CI for  $\Delta$ . Then we can calculate these bounds by considering which particular noncentral  $t$  distributions are compatible with our data.

Finally, having found upper and lower bounds for  $\Delta$ , we divide these by  $\sqrt{n}$  to find the bounds for  $\delta$ .

We start with our observed  $\bar{x}$  and  $s$ , to calculate

$$t_{n-1,\Delta} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \quad [17]$$

which will have a particular noncentrality parameter. Now if we take  $C = 95$ , for the lower bound for  $\Delta$ ,  $\Delta_L$ :

$$\Pr(t_{n-1,\Delta_L} \geq \frac{\bar{x} - \mu_0}{s/\sqrt{n}}) = 0.025 \quad \text{and, for the upper bound for } \Delta, \Delta_U:$$

$$\Pr(t_{n-1,\Delta_U} \leq \frac{\bar{x} - \mu_0}{s/\sqrt{n}}) = 0.025. \quad [18]$$

So  $t_{n-1, \Delta_L}$  is the noncentral  $t$  distribution that gives rise to the observed  $t$  value with a probability of 0.025 in its upper tail, and  $t_{n-1, \Delta_U}$  is the distribution that gives rise to the observed  $t$  value with a probability of 0.025 in its lower tail. There is no formula that can give us these values of  $\Delta$  directly, and statistical software typically uses an iterative algorithmic search to find the values (cf. Smithson, 2001).

Once the upper and lower bounds for  $\Delta$  are found, it is easy to apply [16] to compute the CI for  $\delta$ . It is  $(\Delta_L / \sqrt{n}, \Delta_U / \sqrt{n})$ .

Informally, we described Method 2 as one way to find the CI for  $\mu$  by sliding the two curves in Figure 2 until the upper tail of the lower distribution, and the lower tail of the upper distribution—tails defined relative to the position of our sample mean—each was of size  $\frac{1}{2}(100-C)/100$ . We can use an identical rationale to find the CI for  $\Delta$ , the only difficulty being that the shape of the distribution changes as we ‘slide’ it along the  $t$  axis, because  $\Delta$  itself is one parameter determining the shape.

Figure 6 shows a combination of the data presentation of Figure 4, including the axis for  $d$ , and the two curves of Figure 2, but here they are noncentral  $t$  distributions. The  $\Delta$  parameters for these two curves define the bounds of the CI for  $\Delta$ , then the CI for  $\delta$  is easily calculated using [16].

#### *ESCI for the confidence interval for $\delta$*

In **CIdelta**, from which Figure 6 is derived, the user can slide the two distributions until the tail sizes are as desired, then read off the CI bounds for both  $\Delta$  and  $\delta$ . Alternatively you can click a button to have Excel do the work, and simply watch the curves move until the correct tail sizes are achieved through the iterative estimation process. You can enter

your own data, and find the CI for  $\mu$ —which as usual is in original units—and the CI for the standardised effect size, Cohen’s  $\delta$ , relative to your chosen reference value  $\mu_0$ . The CI bounds for  $\mu$  are read off the upper, original score axis, and the CI bounds for  $\delta$  are read from the lower  $d$  axis. For our example the CI for  $\mu$  is (28.4, 36.9) and the CI for  $\delta$  is (0.39, 1.86).

As well as the interactive diagrams, **CIdelta** provides numerical values of descriptive statistics, CIs and  $t$  and  $p$  values. You can also mark the endpoints of the two CIs calculated from your data, and note where these fall on both the original score axis, and the  $d$  axis. In our example, verticals through the endpoints of the CI for  $\delta$  intersect the upper, original score axis at 27.6 and 37.5.

Consideration of the derivations above, and exploration with **CIdelta** can lead to a number of conclusions, including:

- These are different CIs, for different parameters, on different axes.
- The CI for  $\mu$  is, as noted earlier, symmetric and centred on the observed  $\bar{x}$ .
- The CI for  $\delta$  is in general not symmetric around the mean  $d$ .
- It is important to think of  $d$  and  $\delta$  as measuring a standardised distance from a particular chosen  $\mu_0$ . So if you change  $\mu_0$  there will be *no* change in the CI for  $\mu$ , but the CI for  $\delta$  may change considerably.
- If in **CIdelta** you change  $\mu_0$ , and observe the CI for  $\delta$  changing, sometimes it—the lower CI in Figure 6 or on screen—will appear shorter than the upper CI and sometimes longer. If  $\mu_0$  is positioned within the CI for  $\mu$ , the lower CI is typically shorter on screen than the upper. However as  $\mu_0$  is moved outside the upper CI,

especially if it is moved far outside, the lower CI is typically longer on screen and can be considerably longer.

- At very small sample sizes, and thus small  $df$ , the CI for  $\delta$  is particularly wide, and the tendencies noted above are particularly marked.
- Presenting the two CIs on a single diagram (Figure 6 and **CIdelta**), and making length comparisons as above, is misleading in the sense that it may suggest a fixed mapping between the two scales on which the two CIs are represented. In fact the  $d$  scale not only depends on the placing of  $\mu_0$ , which sets the zero, but also on the value of  $s$  for this particular sample, which sets the unit of the lower axis. If a different independent sample were taken we would expect the two CIs to be different but, in addition, we would expect  $s$  to be different and so even the scale on which the CI for  $\delta$  is represented would be different.

### **Confidence intervals for $\mu$ and $\delta$**

Comparison of the CIs for  $\mu$  and  $\delta$  may prompt a number of thoughts. First, the two intervals are calculated from the same data: Should one be preferred? The answer is that both are legitimate and useful; they simply tell us about different parameters. If we are interested in  $\mu$ , the population mean in original units, then the simple CI based on central  $t$  is appropriate. If however we have a quite different aim, and wish to know what the data can tell us about the standardised effect size in the population, relative to our chosen reference value  $\mu_0$ , then the CI for  $\delta$ , based on noncentral  $t$ , is appropriate. We are asking a very different question, so it should not be surprising that the answer we derive from our data is different.

Second, consider statistical significance testing. Earlier we referred to the familiar single sample  $t$  test, based as usual on central  $t$ , which is appropriate for testing a null hypothesis about  $\mu$ . In contrast, testing a null hypothesis about  $\delta$  is not straightforward and requires use of noncentral  $t$ . The most practical way to carry out such a test is to find the CI for  $\delta$ , then to note whether or not this includes the particular  $\delta$  value specified in our null hypothesis.

Thinking further about NHST, examination of Figure 6 may prompt consideration of a vertical line positioned to be captured by the lower but not the upper CI. For example a vertical line through 28.0 on the upper axis would cut the lower axis at 0.45, meaning that  $\mu = 28.0$  corresponds to  $\delta = 0.45$ . (Note that this ‘correspondence’ holds just for the particular  $s$  for our sample, which sets the units of the  $d$  axis.) Surely, we might wonder, if we used the upper CI to test a null hypothesis of  $\mu = 28.0$  and the lower to test the null hypothesis  $\delta = 0.45$ , we would be testing corresponding hypotheses but would come to contradictory decisions? We would reject the null for  $\mu$ , but not for  $\delta$ ? In response, note first that for NHST we need to specify the parameter value in our null hypothesis *in advance*. Second, because the relation between the lower and upper scales depends on  $s$  we cannot expect the null values for  $\mu$  and  $\delta$  we specified in advance to line up vertically: It would be an astonishing coincidence if they did happen to correspond. However even if they did come close to corresponding, and did happen to fall so that just one was captured by the relevant CI, there is no problem. As we keep emphasising, the two CIs are for different parameters and need to be referred to different scales. Just as it should not be surprising that the two intervals appear different in Figure 6, it raises no logical problem if NHST gives different outcomes in the two cases.

If we choose to test  $H_0: \delta_0 = 0$ , as we more commonly might, we are testing the null hypothesis of a zero effect. In this case there can be no difference between two test outcomes. By the definition of  $\delta$  (see [9]), testing  $H_0: \delta_0 = 0$  is equivalent to testing  $H_0: \mu = \mu_0$ , which here is 25. As noted earlier, the noncentral  $t$  distribution with  $\delta = 0$ , and thus  $\Delta = 0$ , is simply central  $t$ , and the test for  $\delta$  is in this case the same as the familiar test for  $\mu$ . Some thought about Method 2 and Figure 6 supports the conclusion that the  $\mu_0$  line in **CIdelta** can never be positioned to be captured by one CI but not the other. If it is positioned exactly at one end of the CI for  $\mu$ , the corresponding end of the CI for  $\delta$  will also fall exactly on the  $\mu_0$  line: That is the graphical equivalent of noting that when  $\mu = \mu_0$ ,  $\delta = 0$ .

Finally, consider the Method 1 explanation we gave earlier for the CI for  $\mu$ . An explanation like this is given in numerous textbooks. For the simplest case it starts by assuming some particular value for  $\mu$  and a known population standard deviation,  $\sigma$ . It considers the sampling distribution of the sample mean and notes the half width of the interval, centred on the assumed  $\mu$  value, that contains  $C\%$  of sample means. This half width is then applied either side of the obtained sample mean to give the CI for  $\mu$ . It may seem contradictory that we start with a single particular value for  $\mu$ , but finish with a CI that explicitly countenances the unknown (but fixed)  $\mu$  having any one of a range of values! In fact there is no questionable logic here, just an important step that is easily overlooked: We need to assume that the half width found for our initially chosen  $\mu$  value applies also for any other value of  $\mu$ . Fortunately this is true for the  $\mu$  case. (When we do not assume that  $\sigma$  is known,  $s$  is used as an estimate and the half width of the interval

will vary from sample to sample. However Method 1 can still be used in this case to derive the CI for  $\mu$ , as in [1] to [6] above.) Under our alternative formulation (Method 2, as in Figure 2) once again the half width varies from sample to sample but, most importantly, we do not need to make the assumption that the half width is independent of the value of the parameter being estimated,. This is fortunate because this assumption is not true in the  $\delta$  case.

### **Confidence intervals for cumulation of evidence**

Our Reason 3 for CIs is that they are useful for the cumulation of evidence over studies, and encourage meta-analytic thinking. The case for meta-analysis has been made cogently by, for example, Hunter and Schmidt (1990) and Schmidt (1996), and has been supported by the TFSI (Wilkinson & TFSI, 1999, p. 599).

Schmidt (1996) stated that:

...any single study is rarely adequate by itself to answer a scientific question.

Therefore each study should be considered as a data point to be contributed to a later meta-analysis, and individual studies should be analyzed using not significance tests but point estimates of effect sizes and confidence intervals. (p. 124)

Schmidt argued that CIs provide a good way for evidence to be combined over studies and that they will home in on an accurate value of the parameter of interest, even if individual studies lack precision or give deviant results.

### *Meta-analytic thinking*

As we stated earlier, we believe it is important to promote researchers' use of meta-analytic thinking, and that CIs can be very useful in helping achieve this aim. This would be a substantial conceptual change; as Schmidt (1992) noted: "It [meta-analysis] is a new way of thinking about the meaning of data, requiring that we change our views of the individual empirical study" (p. 1173).

There are several aspects to meta-analytic thinking. First, it should give an accurate and justifiable appreciation of previous research on our question of interest. Second, it should allow us to appreciate our own study as making a probably modest empirical contribution, in the context of that previous research—although the rare individual study can of course change thinking. Third, it encourages us to present our results in a way that makes it easy for future researchers to integrate them into future meta-analyses.

In the social sciences, little work has been done on how CIs can best be used to support such meta-analytic thinking. We suspect there is room for much thought and development. There is also great scope for learning from other disciplines, especially the medical sciences, where general use of CIs (e.g., Altman, 2000) and meta-analysis based on CIs is well established (e.g., Chambers & Lau, 1993; Duval & Tweedie, 2000; Greenland, 1987). We are especially interested in graphical approaches: Light, Singer and Willett (1994), and Altman, discussed graphical ways to present a set of CIs. We offer here a simple way to view and combine a set of CIs for a single parameter.

*Meta-analysis in original units*

Consider a small set of Case 1 studies, and suppose these constitute the previous research on a question of interest. If the same measurement scale was used in each case, and we consider the studies to be sufficiently comparable, the original units means may be combined by simple pooling, as in the ‘bare bones’ fixed effect meta-analysis of Hunter and Schmidt (1990). Based on the whole set of data, the overall simple weighted mean can be calculated, as well as the CI for the assumed common  $\mu$ . If we then carry out our own study, using the same measurement scale, we can include our own data and carry out a second slightly larger meta-analysis. Software is available to assist, for example *MetaWin* (Rosenberg, Adams, & Gurevitch, 2000).

The original units sheet of **MAtHinking** also supports such an analysis and displays the CIs for  $\mu$ , derived from each individual study, and from the two sets of studies (the previous research, and that augmented by our study). We simply enter the means, standard deviations and sample sizes for the previous studies, and for our own study. It is possible also to show the results of NHST (simple  $t$  tests) for each study, although this is not necessary. This allows a comparison of the classical review method of counting statistically significant results and the (much preferable: Schmidt, 1992, 1996) meta-analytic approach of calculating a combined estimate.

*Meta-analysis in standardised units*

**MAtHinking** for original units requires that all studies use the same measurement scale, and dependent variables that are sufficiently comparable for the original scores to be justifiably combined over studies. One of the advantages of  $d$  is of course that the first assumption, use of the same measurement scale, is not necessary. Figure 7, from the

standardised units sheet of **MAtHinking**, shows a display combining the  $d$  values from each of a set of studies, with CI for  $\delta$  shown in each case. The results of our own study are shown, and are integrated into the whole. Finding each CI for  $\delta$  requires iterative calculations based on noncentral  $t$ ; in **MAtHinking** the user needs to click a button to trigger this process. Again NHST results may be displayed if desired.

In combining studies in this simple way we do not mean to dismiss important issues that should be considered in any real meta-analysis. Hunter and Schmidt (1990), for example, discussed at length a considerable number of threats to easy combination of studies, and presented analytic methods that allow the importance of these to be assessed.

Even so, working with **MAtHinking**, in either original or standardised units, may allow a number of useful insights to be gained, including:

- If there are even a few past studies, our own study will have only a modest influence on the final picture, unless our  $n$  is particularly large.
- The most important influences are the magnitude of the effect sizes from the various studies and, especially, the consistency of these over studies.
- The statistical significance status of individual studies means little. Several studies not individually reaching statistical significance can easily give a ‘highly significant’ combined result if the effect sizes are reasonably consistent, at least in direction.
- $n$  is important, but effect size is even more so ( $n$  has influence only via its square root—and to a small extent via  $df$ ).

Whether original or standardised units should be preferred, or both analysed, is a matter for judgement in the particular situation. Using  $d$  may allow cumulation even

when different original units measures have been used in different studies. On the other hand, because  $d$  is influenced by both a mean (the numerator) and a standard deviation (the denominator), effect sizes that are the same in original measurement units but subject to different error—perhaps arising from different measurement error—will have *different* standardised effect sizes.

#### *The diversity of meta-analysis*

Meta-analysis can be appropriate for a wide range of situations and a variety of research goals. For example Rosenthal (1991) started with a meta-analysis of two means and Hunter and Schmidt (1990, chapter 10) considered meta-analytic combination of results within a single study, which may comprise replications with a single dependent variable, or a number of different dependent variables. Closer to the other end of the spectrum was Glass's (1976) classic study of psychotherapy that drew on 375 earlier research studies. The goal of a meta-analysis may be simple quantitative combination of a few effect size estimates, or may be to draw on sophisticated techniques to identify moderator variables and to contribute to theory development (Cook et al., 1992). Rubin (1990) spoke of using meta-analysis "for understanding the underlying science" (p. 155). Such different meta-analytic situations are of course likely to raise different issues about the combining of a variety of measures, and the advisability of working in original or standardised units.

Within a general encouragement for meta-analytic thinking we would like to emphasise the value of small-scale meta-analysis in the social sciences to integrate results from a single researcher, a single laboratory or group of collaborators. Small sets of measures from such a source are likely to be closely related conceptually, and less subject to diverse sources of measurement error. Such small-scale meta-analysis should clarify

findings rapidly, and if necessary provide input to later meta-analyses of larger scope.

We hope that **MAtHinking** can illustrate and encourage such manageable, small-scale analyses, while also promoting understanding of some basic meta-analytic ideas.

### **Statistical power**

Informally, power is the chance that, if there is a real effect, our experiment will find it.

More formally, power is the probability that we will reject the null hypothesis, if it is false. Statistical power is thus the conditional probability:

$$\Pr(\text{reject } H_0 \mid H_0 \text{ is false}).$$

If  $H_a$  is the alternative hypothesis, power is:

$$\Pr(\text{reject } H_0 \mid H_a \text{ is true}).$$

To calculate this we need to use a point value for the population parameter specified by  $H_a$ ; we will consider just the case where this is  $\mu_a$ .

Any value for power reflects a number of factors, including population variability, design of the study,  $n$ , the chosen Type 1 error rate ( $\alpha$ ), and  $\mu_a$ . Often it is useful to take a sensitivity approach to the complex interrelation of so many variables by considering, for example, how power varies with  $n$ , or with  $\mu_a$ . Such calculations at the research planning stage can guide the choice of  $n$ . After data are collected power calculations are often used to give further insight into the precision of such an experiment, perhaps now using estimates (e.g., for population variability) based on the data. We will suggest, however, that the precision of an experiment is better described by a CI.

Calculation of power after collecting the data ('post hoc power') was discussed and supported by Cohen (1988) and is a practice of some social scientists. However, the

meaningfulness of such calculations has been questioned by some statisticians: “The arbitrariness of power specification is of course absent once the data are collected, since statistical power refers to the probability of obtaining a particular type of data; it is thus not a property of data sets.” (Greenland, 1988, p. 236)

In other words power, which is a conditional probability that a particular NHST outcome will occur, does not make sense with reference to a particular dataset—for which that NHST outcome either did or did not occur. Power, whether calculated in advance or post hoc, should be considered as a property of a potential experiment having particular features (including design, and sample size) and assuming particular values for population variability and effect size. Again, we emphasise the value of a sensitivity approach.

#### *Power and NHST*

We expect that, as NHST use declines, CIs become more widely used, and more is understood about how CI width can—as we discuss in a following section—be a good index of experimental precision, then need for the concept of power will in many cases decline. Schmidt (1996) pointed out that:

If significance testing is no longer used, then the concept of statistical power has no place and is not meaningful. In particular, there need be no concern with statistical power when point estimates and confidence intervals are used to analyze data in studies and when meta-analysis is used to integrate findings across studies. (p. 124)

We should note however that power is used also in more complex contexts, including multi-parameter inference, in which NHST may not be readily displaced by CIs.

Use of NHST can and should decrease markedly, at least in many simple situations for which CIs provide a ready replacement. It remains to be seen how widely NHST can be displaced, and what uses of it might persist, and so the extent of the future need for power is not yet clear. In the meantime power remains an important concept. If NHST is conducted, power should be considered. In particular, if a null hypothesis is *not* rejected, it is important that power calculations be examined.

There are two additional reasons for our discussion of power here. First, noncentral  $t$  needs to be used to assess power accurately, for any of our three Cases, although researchers and even textbook authors may not always realise this or may, in practice, use simple approximations. Because we are investigating noncentral  $t$  and its properties and uses, it is natural to discuss power as an application of these distributions. Second, developing a better understanding of how CI width is best used as an index of precision may be assisted by an understanding of power.

#### *The power diagram*

Figure 8 is the standard diagram used to explain power in numerous statistics textbooks. The two curves look identical and may be described as the sampling distribution of the test statistic: one of the curves if the null hypothesis is true, and the other curve if the alternative hypothesis is true. If  $\sigma$  is known and a  $z$  test is being used, the two sampling distributions are identically-shaped normal distributions. However in the much more common case of  $\sigma$  unknown, a  $t$  test statistic should be used. In this case the null hypothesis curve will be a central  $t$  distribution, while the alternative distribution is actually a noncentral  $t$  distribution, with  $\Delta$  indexed by the difference between  $\mu_a$  and  $\mu_0$ , the population parameters postulated by the two hypotheses.

Many textbooks present the power diagram with two seemingly identical curves and do not give a full account of what test statistic is being used, and with what assumptions. Few textbooks make the careful distinction between the known  $\sigma$  condition, in which curves with identical shape are correct, and the situation involving  $t$  in which the shape of the curves is different. Hays (1988) and Howell (1997) are examples of texts that make clear that noncentral  $t$  needs to be used if power is to be calculated accurately in this latter, more usual situation. These texts refer the reader to tables, graphs or approximate formulas for the practical estimation of power.

Figure 9 is from *ESCI Power* and illustrates power when a  $t$  test statistic is appropriate, and so distribution of the test statistic under the alternative hypothesis is a noncentral  $t$  distribution. Figure 9 shows the two curves and the value of power.

*The properties of power*

**Power** can be used to explore the relations among  $df$ ,  $\Delta$ ,  $\alpha$ , and power. Power is illustrated as the area under noncentral  $t$ , an accurate power value is shown, and—using the simple relation [16]—the  $\delta$  value corresponding to  $\Delta$  if  $H_a$  is true is also shown.

Conclusions that might be drawn include:

- For the degenerate case of  $\mu_a = \mu_0$ ,  $\Delta = 0$ , the two curves become identical and power =  $\alpha$ .
- For a two-tail test, as is always assumed in this paper and in *ESCI*, power is the sum of two tail areas under the  $H_a$  curve, although in practical situations one of these areas is almost always negligible.
- Power is extremely sensitive to effect size, that is to  $\Delta$ ; the most effective way to have high power is to be trying to find a large effect!

- Power is also sensitive to  $df$ , that is to  $n$ , but less so than to effect size.
- Many of the earlier conclusions about the shape and behaviour of the noncentral  $t$  distributions have implications also for power.

**Power** also gives some findings that may be surprising to many researchers. It shows the percentages of NHST results expected to reach the conventional  $p < 0.05$  (\*),  $p < 0.01$  (\*\*), and  $p < 0.001$  (\*\*\*) levels of statistical significance if  $H_a$  is true. For the  $p < 0.001$  level these expected percentages may seem in many cases surprisingly high. For example, if  $n = 20$  and  $\delta = 0.52$  (a medium-sized effect), the power (for  $\alpha = 0.05$ ) is only 0.60, yet fully 10.2% of experiments (when  $H_a$  is true) will give a  $p < 0.001$  statistical significance test result. So  $10.2/0.60 = 17\%$  of those results that reach conventional significance ( $\alpha = 0.05$ ) will be \*\*\*. However after a \*\*\* result (in fact after any result, still assuming  $H_a$  is true) the chances are  $1 - 0.60 = 0.40$ , fully four in ten, that a replication would not even reach significance,  $p < 0.05$ ! In summary, a coveted \*\*\* result is not rare in many situations of only moderate power, and even then a replication is far from guaranteed to be even \*!

This observation raises interesting unexplored questions for those interested in statistical cognition: Do researchers realise that a \*\*\* result is not rare even for a medium-sized effect and moderate power? Do they realise what low or moderate power implies about replication, however small the  $p$  value obtained in a particular case?

#### *Approximations for power*

It has long been recognised that noncentral  $t$  is the distribution needed for the correct analysis of power in the simple cases we are considering (e.g., Kendall & Stuart, 1961, p. 255). However calculation difficulties have meant that approximations have been widely

used for practical calculations. Cohen (1988) used various approximate methods to calculate many of his power tables and in many of his recommended procedures involving power. His emphasis was on substantive interpretation, and a good appreciation of the relations between power and various other characteristics of an experiment, rather than on precise calculations.

One simple approximation is, as mentioned above, to use two normal distributions and to calculate a power value as if  $\sigma$  is known and a  $z$  rather than a  $t$  test is being conducted. This approximation gives power values that are accurate or close to accurate, to two decimal places, for  $n$  not small (say 20+), and  $\alpha$  not small (say not less than 0.05). The approximate values are if anything overestimates, because no account is taken of the need to estimate  $\sigma$ , and for small  $n$  and small  $\alpha$  the values can be considerably too high.

Now that software, including *SPSS* and *ESCI Power*, is readily available to give accurate calculation of power, it should no longer be necessary to use approximations, at least for the simple situations we discuss here. However we should bear in mind Cohen's (1988) approach: Concern over precision of the calculation method must not blind us to the assumptions we are making. The values we enter for  $\sigma$  and  $\delta$ , for example, may be quite speculative. A sound strategy is to take a sensitivity approach and to examine how power varies with the other variables.

#### *A range of values for power*

To calculate statistical power we need to specify a  $\delta$  value. For a priori power this may be a value chosen as being of practical or theoretical interest, or we may calculate power for one or more of Cohen's conventional sizes of 0.2, 0.5 and 0.8. After conducting the experiment we have the additional option of calculating power for  $\delta$  equal to the observed

*d.* In addition we can calculate the CI for  $\delta$ , based on our data. Every  $\delta$  value in that interval is an effect size value compatible with our data, in the sense discussed earlier. Each of those  $\delta$  values can be used to find a power value, so from our CI for  $\delta$  we can derive a range of values for power. This is simply the set of power values that correspond to the set of  $\delta$  values that is the CI for  $\delta$ .

Note that this set of power values should *not* be regarded as a ‘CI for power’ because a CI is an interval estimate for a parameter that has (in our simple cases) a single true, but unknown, value. Power is best thought of not as such a parameter, but as a descriptive feature of an experiment that has particular characteristics (including  $n$ , design,  $\sigma$  and  $\alpha$ ). Our focus should be on how power varies with such characteristics and not on a quest for a supposed single ‘true’ value.

In addition to giving the CI for  $\delta$ , **CIdelta** gives the corresponding range of power values. Exploration is likely to lead to the conclusion that, for a broad range of realistic datasets, this range is surprisingly wide. Thus in many cases our results may reasonably have arisen from a small population effect size and low power, or a large effect size and power near or equal to 1.0. This uncertainty may increase the attractiveness of CI width as an index of the precision of an experiment.

### **Confidence intervals and precision**

Jacob Cohen, a highly influential reformer, for more than two decades worked hard to encourage psychologists to calculate and think about statistical power (e.g. Cohen, 1990, 1994). His book (Cohen, 1969, 1988) introduced Cohen’s  $\delta$ , which was important for the development of meta-analysis; it remains a classic reference for power analysis. Cohen

argued that psychologists seldom realised how low their statistical power usually was, and so NHST results were misinterpreted and much research effort was wasted. If power were routinely estimated and considered at the research planning stage, and routinely reported as part of the analysis of results, psychological research could be much improved.

In 1994 N. R. Thomason asked Cohen why he had chosen statistical power, rather than the use of CIs, as the focus of his reform efforts, and why he had persisted with power for so long. Thomason (personal communication from N. R. Thomason, 26 February 2001) reported Cohen's response by saying that

only someone who was into the NHST mindset could fully understand why he pressed on with power for so many years. It is a way of thinking that is so conceptually attractive that it took him years to free himself from it. However, he had come to believe that adopting CIs is a higher priority than increased use of power and that it would have been better if he had from the start chosen CIs as the focus of his efforts.

CI width reflects a number of aspects of the precision of a study, including the amount of variability in the population, the sample size and thus sampling error, and the amount of error in the dependent variable. Statistical power is also influenced by all these factors, but is defined relative to an effect size that is set by the value of the population parameter specified by the alternative hypothesis. As noted earlier, power is strongly influenced by this effect size. For example a wide CI might be associated with high power simply because the chosen effect size, set by the alternative hypothesis, is

very large. By contrast, effect size has no systematic influence on the width of the original units CI.

Power is defined in terms of, and cannot be divorced from NHST. By being tied to a particular value of the population parameter—a particular effect size—and given the desirability of de-emphasising NHST, power has distinct disadvantages. CI width is likely to become recognised by social scientists as a more generally useful guide to the precision of studies with simple designs like those we discuss in this paper. Steiger and Fouladi (1997) supported use of CI width as an index of precision, but cautioned that “the width of a confidence interval is generally a random variable, subject to sampling fluctuations of its own, and may be too unreliable at small sample sizes to be useful for some purposes” (p. 254).

Our Excel simulations provide a number of ways of exploring the relations between CI width and important features of an experiment, including population variance, sample size, design (our three Cases), and whether original or standardised units are the focus. We hope these can help develop appreciation of CI width as a broadly useful indicator of precision.

## CASE 2: TWO INDEPENDENT GROUPS

In Case 2 we have two independent random samples from normal populations, of size  $n_1$  and  $n_2$ , and are interested in the difference between the group means. We need give only

an outline presentation here because the rationale is closely analogous to that for Case 1.

It is worth comparing the formulas given here with the corresponding Case 1 formulas.

### Case 2, original units

Consider, for original scores, that

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_{01} - \mu_{02})}{S\sqrt{(1/n_1 + 1/n_2)}} \sim t_{n_1 + n_2 - 2} \quad [19, \text{cf. } 1]$$

where  $\bar{X}_1$  and  $\bar{X}_2$  are the sample means,  $S$  is an estimate of  $\sigma$  pooled from the two groups (see below), and  $(\mu_{01} - \mu_{02})$  is a reference value for the difference between the population means. In almost all cases this reference value is set to zero; this corresponds to testing a null hypothesis of no difference between the two population means. In our discussion of Case 2 we will consider only the situation where  $(\mu_{01} - \mu_{02}) = 0$ .

Now  $\sigma$  is the population standard deviation, which we assume to be the same in the two populations. The pooled estimate of  $\sigma$  is may be calculated from:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} \quad [20]$$

where  $s_1$  and  $s_2$  are the sample standard deviations (with  $n_1 - 1$  and  $n_2 - 1$  in their respective denominators). The standard error of the difference between the means is  $s\sqrt{1/n_1 + 1/n_2}$ , which becomes  $s\sqrt{2/n}$  when  $n_1 = n_2$ .

The original units 95% CI for the difference between the means may be calculated as:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1 + n_2 - 2}(0.975) \times s\sqrt{1/n_1 + 1/n_2} . \quad [21, \text{cf. } 6]$$

### *Comparisons of Cases 1, 2 and 3*

Cases 1, 2 and 3 involve the estimation of different parameters and so are not directly comparable, but they are sufficiently analogous for comparisons of some quantities to be of interest. Comparison of Cases 2 and 3 (see below) are of most practical interest because in many situations either might be used and the researcher needs to decide between them.

For Case 2 with equal-sized groups (and using  $n = n_1 = n_2$ ) the standard error of the difference between the means is  $s\sqrt{2/n}$ . Other things being the same, this is  $\sqrt{2}$  or 1.41 times as large as the standard error of the single mean in Case 1. Consequently the Case 2 CI is wider than the Case 1 CI by a similar factor, although the larger  $df$  in Case 2 has an influence—usually small—in the opposite direction. This comparison is, of course, most meaningful when the Case 1 data are a set of difference scores.

### **An original units example, Case 2**

Figure 10 shows dot plots of the data points in two independent groups, the sample means and their individual CIs, and the CI for the population mean of the difference between the means. (You can use the Case 2 sheet in **CIoriginal** to make these calculations and generate a similar display for your own data.) Variations on Figure 10 (or explorations with **CIoriginal**) illustrate some basic features of the original units CI for the population mean, including:

- The CI is centred on the sample difference between the means and is symmetric.
- Higher chosen  $C$  requires a wider CI.
- Larger sample sizes give a shorter CI.

- CIs for the difference between the means in Case 2 are generally wider than the CI for the mean of either of the groups.
- The Case 1 conclusion that many realistic datasets in the social sciences give 95% CIs that are disappointingly wide applies also to Case 2.

### Case 2, standardised units

The population standardised effect size can be defined as:

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \quad [22, \text{cf. } 9]$$

where  $\mu_1$  and  $\mu_2$  are the two population means. (Recall that we are assuming throughout the Case 2 discussion that  $(\mu_{01} - \mu_{02})$ , the comparison value for the difference between the population means, is zero.)

The sample effect size can be calculated as:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s} \quad [23, \text{cf. } 10]$$

where  $s$  is an estimate of the pooled population standard deviation, calculated using [20].

The conventional  $t$  statistic for testing the hypothesis that the difference between the population means is zero is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{1/n_1 + 1/n_2}}. \quad [24, \text{cf. } 11]$$

Combination of the last two formulas gives:

$$d = t\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad [25, \text{cf. } 12]$$

which becomes  $d = t\sqrt{2/n}$  when the two samples are of equal size. Again this is a simple and appealing relation that is worth remembering.

The noncentrality parameter for the distribution of  $t$  is:

$$\Delta = \frac{\mu_1 - \mu_2}{\sigma\sqrt{1/n_1 + 1/n_2}}. \quad [26, \text{cf. } 15]$$

Therefore as in the one sample case there is a simple relation between  $\delta$  and  $\Delta$ :

$$\delta = \Delta\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \quad [27, \text{cf. } 16]$$

The Method 2 argument for finding a CI for  $\delta$  for Case 2 is essentially the same as that for Case 1. Use the Case 2 sheet of **CIdelta** to see two noncentral  $t$  distributions move until the tail conditions (cf. [18] and Figure 6) are satisfied and the  $\Delta$  values, and thus the  $\delta$  values, are found for the lower and upper ends of the CI.

Exploration with **CIdelta** leads to conclusions that are very similar to those for Case 1 (a single group), but with the CIs in Case 2 (two independent groups) being as we would expect generally wider than those in Case 1.

### CASE 3: SINGLE GROUP, REPEATED MEASURE

In Case 3 we have two measures from each of  $n$  experimental units and are interested in the mean of the  $n$  differences, more specifically how far this mean is from some reference value, most commonly zero. We may have a pre- and post-measure from each of  $n$  participants, or the experimental units may be pairs of participants who are matched in

some natural way (e.g., parent and child), or that have been assigned to a pair because for example they have similar scores on some relevant matching variable or variables. We assume the difference scores to be a random sample from a normal population.

### **Case 3, original units**

In original units the effect size is simply the mean of the  $n$  differences. In fact for original units Case 3 reduces completely to Case 1 by the simple expedient of taking the difference scores as the data. All the discussion and original score formulas given above for Case 1 are applicable: Simply apply them to the differences. All the simulations and conclusions based on the *ESCI* tools for original scores also apply. The Case 3 sheet in **CIoriginal** calculates the differences but otherwise is essentially the same as the Case 1 sheet.

To the extent that matching or the use of repeated measures leads to the two measures being positively correlated, Case 3 is a sensitive design. This sensitivity arises because the width of the original score CI for mean difference—the original score effect size—is based on the variability of the *differences*, and not on the variability of the  $n$  scores on either measure. To the extent the measures are positively correlated, the variability of differences is smaller than this latter variability. Other things being equal the CI will be narrower for Case 3 than for Cases 1 or 2.

We noted earlier that the CI for Case 2 is wider than the CI for Case 1 by a factor of approximately  $\sqrt{2}$ . Therefore, because Case 3 reduces to a Case 1 sample of difference scores, the CI for Case 3 will at worst be narrower than the CI for Case 2 by a factor of approximately  $\sqrt{2}$ , other things being equal. To the extent that the measures in

Case 3 are positively correlated, the Case 3 CI width will be narrower still. This is an important consideration when choosing between Case 2 and 3 designs, in situations when either may be used.

### **Case 3, standardised units**

Unfortunately the situation is more complicated when we turn to standardised units. Any standardisation requires choice of an appropriate standard deviation to use in the denominator as the reference unit. For Case 3, we could standardise the mean difference by dividing it by  $s_D$ , the standard deviation of differences, or by  $s$ , the standard deviation of the measures themselves. (We assume the population variance to be the same for each measure, just as in Case 2 we assumed a common  $\sigma$  for the two populations.) We could analyse both ways, if we judged this appropriate. Glass, McGaw, and Smith (1981, pp. 116-117), among others, made clear that the choice should be based on substantive considerations in the application area. They noted that the choice is crucial because  $s_D$  is often in practice considerably smaller than  $s$ , and so a standardised effect size based on  $s_D$  will be considerably larger than one based on  $s$ .

Cohen (1988, pp. 48-49) discussed this issue and noted that “If the investigator is content” (p. 48) to choose  $s_D$ , then Case 3 reduces fully to Case 1 for standardised units as well as for original units. However Cohen expressed a general preference for using  $s$  as the basis for standardisation.

Note that the question of standardisation of the effect size is different from choice of denominator for a statistical significance test. Whichever standard deviation is used in

the estimate of  $\delta$ , it is appropriate to use  $s_D$  in the denominator of the conventional paired  $t$  test of the hypothesis of zero mean difference.

*Standardisation using  $s_D$*

A researcher who decides, for a particular Case 3 example, that standardisation using  $s_D$  is appropriate is judging that the mean difference can best be assessed in terms of the variability of difference scores. Such an example would need no further comment because all the discussion, formulas and *ESCI* simulations given earlier for Case 1 would apply, for standardised units as well as original units. These could all simply be applied to the differences.

*Standardisation using  $s$*

It is easy to find examples that align with Cohen's (1988, p. 49) preference in that standardisation against  $\sigma$  (or  $s$ , our best estimate of  $\sigma$ ) is appropriate. Suppose we use a verbal ability test known to have  $\mu = 100$  and  $\sigma = 15$  (so here we use  $\sigma$  itself, rather than an estimate) in the relevant population, and we investigate the effect of having a healthy breakfast on test scores. A group of  $n$  children are tested on parallel forms of the test, each child being tested on different days with and without a healthy breakfast—no doubt we would counterbalance the order of testing. We calculate our mean difference to be 4.1 points in favour of the healthy breakfast. It would seem most useful to assess the mean difference against  $\sigma$  and to report an advantage of  $d = 4.1/15 = 0.27$  (in Cohen's terms, a small effect). The calculated standard deviation of difference scores may be relatively small, say 7.7, and this value would be used, quite correctly, to calculate the original score CI for the effect (and the  $t$  value if we wished to conduct a significance test). However using this value to standardise the effect size (which gives  $4.1/7.7 = 0.53$ )

probably does not make substantive sense: Our natural reference for thinking about verbal ability scores is  $\sigma = 15$ , and not the variability in differences.

Choosing  $\sigma$  for standardisation, as is appropriate in at least many Case 3 situations, we can define the standardised effect size as

$$\delta = \frac{\mu_D}{\sigma} \quad [28, \text{cf. } 9, 22]$$

and can calculate the sample mean difference by using

$$d = \frac{\bar{x}_D}{s} \quad [29, \text{cf. } 10, 23]$$

where  $\mu_D$  is the population mean difference,  $\sigma$  is the population variance of the measures, assumed the same for each, and  $\bar{x}_D$  is the mean difference calculated from the data. The denominator  $s$ , our best estimate of  $\sigma$ , may be the standard deviation of the  $n$  pre-measures or, probably better, the standard deviation of the  $n$  means of pre- and post-measures for each pair.

The major difficulty with the Case 3 effect size standardised against  $\sigma$  now becomes apparent. The numerator  $\bar{x}_D$  is normally distributed with standard deviation  $\sigma_D/\sqrt{n}$  whereas the denominator  $s$  is closely related to a  $\chi^2$  distribution with the involvement of  $\sigma$ . Because  $\sigma_D$  is in general not equal to  $\sigma$ ,  $d$  does not follow a  $t$  distribution, central or noncentral. Because the distribution of  $d$  is so complex, we cannot provide a formula or simulation for determining an accurate CI for this standardised effect size for Case 3.

One approach to the problem is based on the relation (Hays, 1988, pp. 313-314)

$$\sigma_D^2 = 2\sigma^2(1 - \rho) \quad [30]$$

where  $\rho$  is the population correlation between the two measures. As we would expect,  $\rho = 0$  reduces to Case 2. If  $\rho$  is large then  $\sigma_D$  is small and so the CI is short and Case 3 is, again as we would expect, especially advantageous. Both Cohen (1988, pp. 48-49) and Glass et al. (1981, pp. 116-118) suggested use of this relation to transform an estimate of  $\sigma^2$  to an estimate of  $\sigma_D^2$ , and so in effect to reduce Case 3 to Case 1. Because in practice that would almost always require estimation of  $\rho$  from the data, and thus involvement of a further source of sampling error, we do not pursue this approach further.

## OVERVIEW AND CONCLUSIONS

Two important aspects of reform of statistical practice in the social sciences are (i) increased use of CIs, and (ii) increased use of effect size measures. It is important to combine the two and consider CIs for effect sizes. However these have received relatively little attention, at least in the case of the simple standardised effect size measure, Cohen's  $\delta$ . One reason for this is, no doubt, that CIs for  $\delta$  involve noncentral  $t$  distributions, which have been little known to many social science researchers.

It is important for statistical reform that techniques for calculating and reporting CIs for effect size measures are widely available and understood. As a contribution to this goal we have presented a discussion centred on CIs for Cohen's  $\delta$ . We considered four general ways that CIs are valuable:

- They give point and interval information that is easily understood, and therefore they facilitate substantive interpretation.

- The link between CIs and NHST can help the learning and understanding of both.
- CIs support meta-analysis, and meta-analytic thinking. Promoting these is a further important aim of reform.
- CI width gives information about precision that may be more useful and accessible than a statistical power value.

We used Case 1, the single sample design, as the context for most of the discussion, and considered measures in original units as well as the standardised measure Cohen's  $\delta$ . In order to discuss CIs for  $\delta$  we needed to describe noncentral  $t$  distributions and their properties. Then we extended the discussion beyond CIs to include simple meta-analysis and statistical power. We gave briefer discussions of Case 2, the two independent samples design, and Case 3, the single sample, repeated measures design.

A key factor is that  $\delta$  involves a distance measure (mean or mean difference) in the numerator and a standard deviation in the denominator, and in practice both these quantities need to be estimated from the data. This is the reason noncentral  $t$  distributions are required. Investigation of the CI for  $\mu$  and CI for  $\delta$  led to a number of conclusions, including:

- These two CIs are calculated from the same data, but are quite different: They estimate different parameters and are expressed on different scales. Judgement is needed to guide the choice to use one, the other, or both, in a particular situation.
- Calculating the CI for  $\delta$  is not straightforward, and usually relies on software that applies an iterative algorithm to find the ends of the CI.

- The CI for  $\delta$  requires specification of a comparison value—e.g.,  $\mu_0$  for Case 1—and the value chosen has a large influence on the CI for  $\delta$ , but no influence on the CI for  $\mu$ .
- Noncentral  $t$  is needed if calculation of power is to be accurate, unless  $\sigma$  is known.
- In some circumstances when there is a real effect but power is relatively low, power calculations show that ‘highly statistically significant’ ( $P < .001$ ) results will not be rare. Some researchers may find this surprising.
- Simple meta-analysis can be carried out using either original or standardised units. Effect sizes are the most influential factor.
- Case 2, the two independent samples design, gives conclusions similar to those for Case 1, but other things being equal CIs are generally wider.
- Case 3, the single group repeated measures design, is attractive where it is applicable because it can give short CIs. The higher the correlation between the two measures, the shorter the CI for the mean of the differences.
- In Case 3  $\delta$  may be based on standardisation by either  $\sigma_D$  or  $\sigma$ . If  $\sigma_D$  is judged appropriate, Case 3 reduces simply to Case 1, but if  $\sigma$  is considered appropriate, the sampling distribution of  $d$  is complex, and we do not present an exact procedure for calculating the CI for  $\delta$ .

We illustrate the discussion throughout by reference to the *ESCI* software, which can be used to calculate CIs in original units and CIs for  $\delta$ , as well as power and basic meta-analysis, for a number of simple situations. We hope especially that explorations using this software will assist researchers and students come to understand better the

concepts touched on in this paper. Many of these are at the heart of reform of statistical practice, which currently needs to be a central concern across the social sciences.

## References

Altman, D. (2000). Clinical trials and meta-analyses. In D. Altman, D. Machin, T. Bryant & M. Gardner (Eds.) Statistics with confidence (2<sup>nd</sup> ed.) (pp. 120-138). London: BMJ Books.

American Psychological Association (1994). Publication Manual of the American Psychological Association (4<sup>th</sup> edition). Washington, DC: Author.

Bakan, D. (1966). The test of significance in psychological research. Psychological Bulletin, *66*, 423-437.

Chambers, T., & Lau, J. (1993). Meta-analytic stimulus for changes in clinical trials. Statistical Methods in Medical Research, *2*, 161-172.

Chandler, R. (1957). The statistical concepts of confidence and significance. Psychological Bulletin, *54*(5), 429-430.

Cohen, J. (1969). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2<sup>nd</sup> edition). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, *45*, 1304-1312.

Cohen, J. (1994). The earth is round ( $p < 0.05$ ). American Psychologist, *49*, 997-1003.

Cook, T., Cooper, H., Cordray, D., Hartmann, H., Hedges, L., Light, R., Louis, T., & Mosteller, F. (1992). Meta-analysis for explanation: A casebook. New York: Russell Sage Foundation.

Cowles, M. (1989). Statistics in psychology: An historical perspective. Hillsdale, NJ: Erlbaum.

Cox, D., & Hinkley, D. (1974). Theoretical statistics. London: Chapman & Hall.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel plot based method of testing and adjusting for publication bias in meta-analysis. Biometrics, *56*, 276-284.

Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. Educational & Psychological Measurement, \*\*\*\*\*

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the Journal of Applied Psychology: Little evidence of reform. Educational & Psychological Measurement, *61*, 181-210.

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., Anderson, J., & Goodman, O. (2001). Reform of statistical inference in psychology: The case of Memory and Cognition. Manuscript in preparation.

Finch, S., Thomason, N., & Cumming, G. (2001). Past and future APA guidelines for statistical practice. Theory & Psychology, \*\*\*\*\*

Gigerenzer, G., & Murray, D. (1987). Cognition as intuitive statistics. Hillsdale, NJ: Erlbaum.

Glass, G. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, *5*, 3-8.

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

- Grant, D. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. Psychological Review, 60, 54-61.
- Greenland, S. (1987). Quantitative methods in the review of epidemiologic literature. Epidemiologic Reviews, 9, 1-30.
- Greenland, S. (1988). On sample-size and power calculations for studies using confidence intervals. American Journal of Epidemiology, 128, 231-237.
- Hays, W. (1988). Statistics (4<sup>th</sup> ed.). NY: Holt, Rinehart, Winston.
- Hogben, D., Pinkham, R., & Wilk, M. (1961). The moments of the non-central  $t$ -distribution. Biometrika, 48, 465-468.
- Howell, D. (1997). Statistical methods for psychology (4<sup>th</sup> ed.). Belmont, CA: Duxbury.
- Hunter, J.E. (1997). Needed: A ban on the significance test. Psychological Science, 8, 3-7.
- Hunter, J., & Schmidt, F. (1990). Methods of meta-analysis: Correcting error and bias in research findings. Newbury Park, CA: Sage.
- Hyde, J.S. (2001). Reporting effect sizes: The roles of editors, textbook authors, and publication manuals. Educational and Psychological Measurement, 61, 225-228.
- Kendall, M., & Stuart, A. (1961). The advanced theory of statistics (Vol. 2). London: Charles Griffin.
- Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. Educational and Psychological Measurement, 61, 213-218.
- LaForge, R. (1967). Confidence intervals or tests of significance in scientific research. Psychological Bulletin, 64, 446-447.

Light, R., Singer, J., & Willett, J. (1994). The visual presentation and interpretation of meta-analysis. In H. Cooper & L. Hedges (Eds.), The handbook of research synthesis (pp. 439-453). New York: Russell Sage Foundation.

Lockhart, R. (1998). Introduction to statistics and data analysis for the behavioral sciences. New York: W. H. Freeman.

Loftus, G. (1993). A picture is worth a thousand p values: On the irrelevance of hypothesis testing in the microcomputer age. Behavior Research Methods, Instruments & Computers, *25*, 250-256.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting & Clinical Psychology, *4*, 806-843.

Moore, D., & McCabe, G. (1993). Introduction to the practice of statistics (2<sup>nd</sup> edition). New York: Freeman.

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

Owen, D. (1968). A survey of properties and applications of the noncentral *t*-distribution. Technometrics, *10*, 445-478.

Pearson, E., & Hartley, H. (1972). Biometrika tables for statisticians, Vol. 2. Cambridge: Cambridge University Press.

Rosenberg, M., Adams, D., & Gurevitch, J. (2000). MetaWin: Statistical software for Meta-Analysis. Sunderland, MA: Sinauer.

Rosenthal, R. (1991). Meta-analytic procedures for social research, Revised ed. Newbury Park, CA: Sage.

Rubin, D. (1990). A new perspective. In K. Wachter & M. Straf (Eds.), The future of meta-analysis (pp. 155-165). New York: Russell Sage Foundation.

Schmidt, F. (1992). What do data really mean? Research finding, meta-analysis, and cumulative knowledge in psychology. American Psychologist, 47, 1173-1181.

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

Schmidt, F., & Hunter, J. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. Harlow, S. Muliak & J. Steiger (Eds.), What if there were no significance tests? (pp. 37-64). Mahwah, NJ: Erlbaum.

Smithson, M. (2000). Statistics with confidence. London: Sage.

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

Steiger, J., & Fouladi, T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. Harlow, S. Muliak & J. Steiger (Eds.), What if there were no significance tests? (pp. 221-257). Mahwah, NJ: Erlbaum.

Thomason, N., Cumming, G., & Zangari, M. (1994). Understanding central concepts of statistics and experimental design in the social sciences. In K. Beattie, C. McNaught & S. Wills (Eds.), Interactive multimedia in university education: Designing for change in teaching & learning (pp. 59-81). Amsterdam: Elsevier.

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. American Psychologist, 53, 799-800.

- Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. Psychological Bulletin, 76, 105-110.
- Wilkinson, L., & Task Force on Statistical Inference, APA Board of Scientific Affairs (1999). Statistical methods in psychology journals: guidelines and explanations. American Psychologist, 54, 594-604.

## Footnotes

<sup>1</sup>References in bold like this are to components of *ESCI* (*Exploratory software for confidence intervals*, pronounced “esky”), our set of tools that run under Microsoft Excel. These provide illustrations of concepts central to this paper, interactivity that allows user exploration, and in some cases the facility to enter your own data and to calculate and display confidence intervals on these data. To obtain *ESCI*, see: [www.psy.latrobe.edu.au/esci](http://www.psy.latrobe.edu.au/esci).

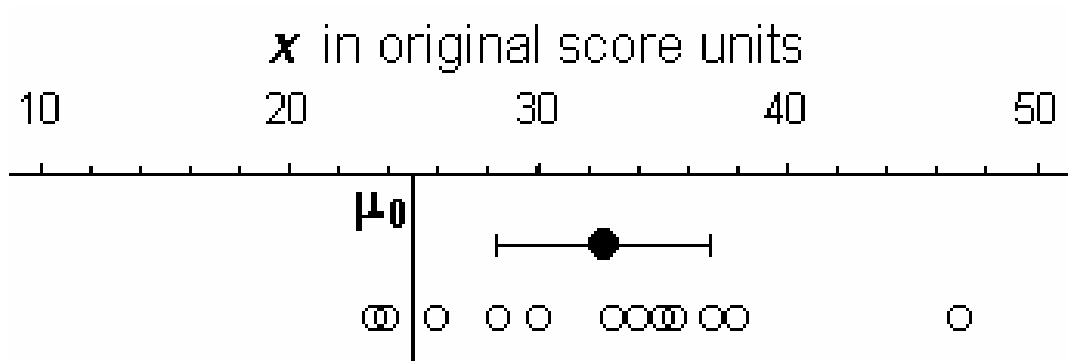
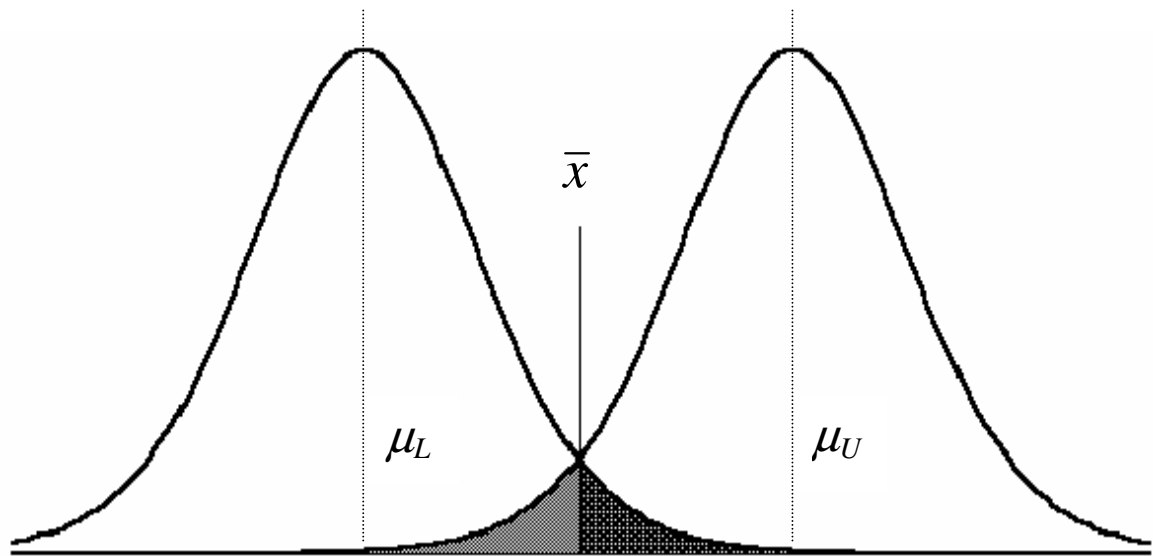


Figure 1. A part image from **CIoriginal**<sup>1</sup>. Twelve data points are shown as a dot plot. The filled dot is the sample mean, 32.6, and the bars mark the 95% CI for  $\mu$ . The vertical line marks  $\mu_0 = 25$ , a value of  $\mu$  chosen for reference. Many numerical results are also reported by **CIoriginal**, including the result of a  $t$  test of the hypothesis that  $\mu = \mu_0$ .



*Figure 2.* The two curves are sampling distributions, with the shape of  $t$  distributions ( $df = 13$ ), scaled to the original units axis. In accord with Method 2 (see text) they have been positioned so that the size of the right tail (dark shading) of the lower distribution that falls to the right of  $\bar{x}$  is  $\frac{1}{2}(100 - C)/100$ , as is the size of the left tail (light shading) of the upper curve that falls to the left of  $\bar{x}$ . ( $C$  is the percent confidence of the CI being constructed.) Then  $\mu_L$  and  $\mu_U$  are the means of the curves positioned in this way, and define the lower and upper ends of the CI for  $\mu$ .

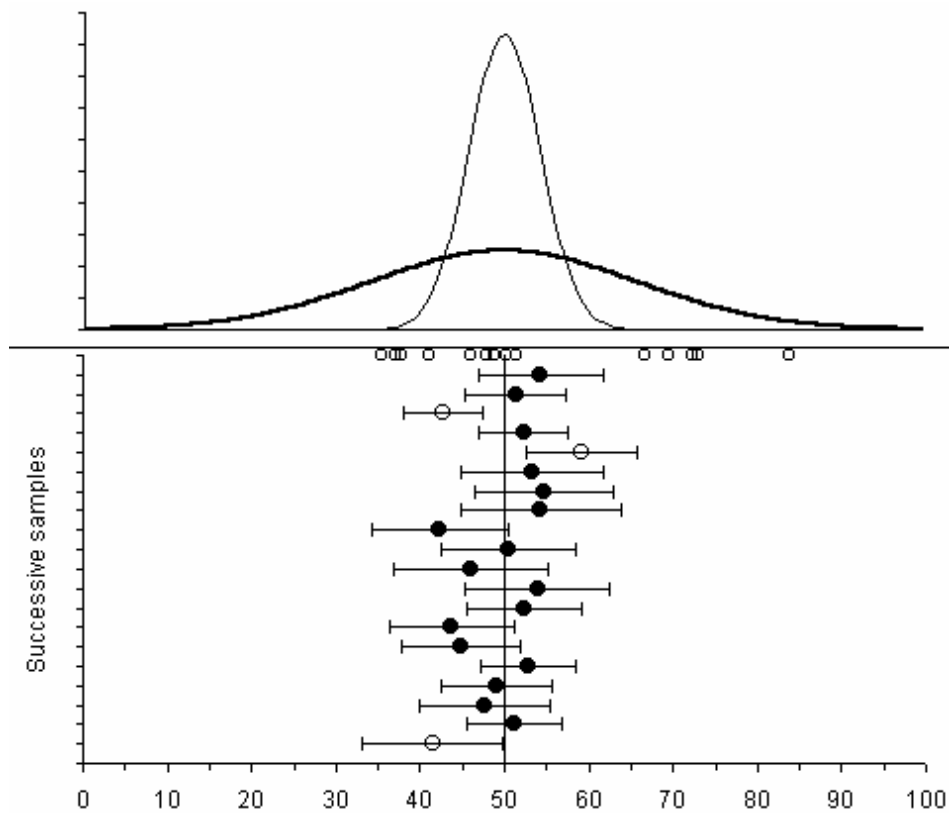


Figure 3. A part image from **CIjumping**<sup>1</sup>. The original units axis is shown below. The heavy curve at the top is the assumed normal population distribution, and the light curve is the sampling distribution of sample means of size  $n = 14$ . The most recent sample, of this size, is shown as a dot plot. Immediately below is the mean and 90% CI for  $\mu$  calculated from this sample. Below that are the means of the previous 19 independent random samples, each shown with its 90% CI. The unfilled dots signal samples for which the CI does not capture  $\mu_0$ , which is here chosen to equal  $\mu$ . Note the haphazard variation from sample to sample in the position of the mean, and in the length of the CI.

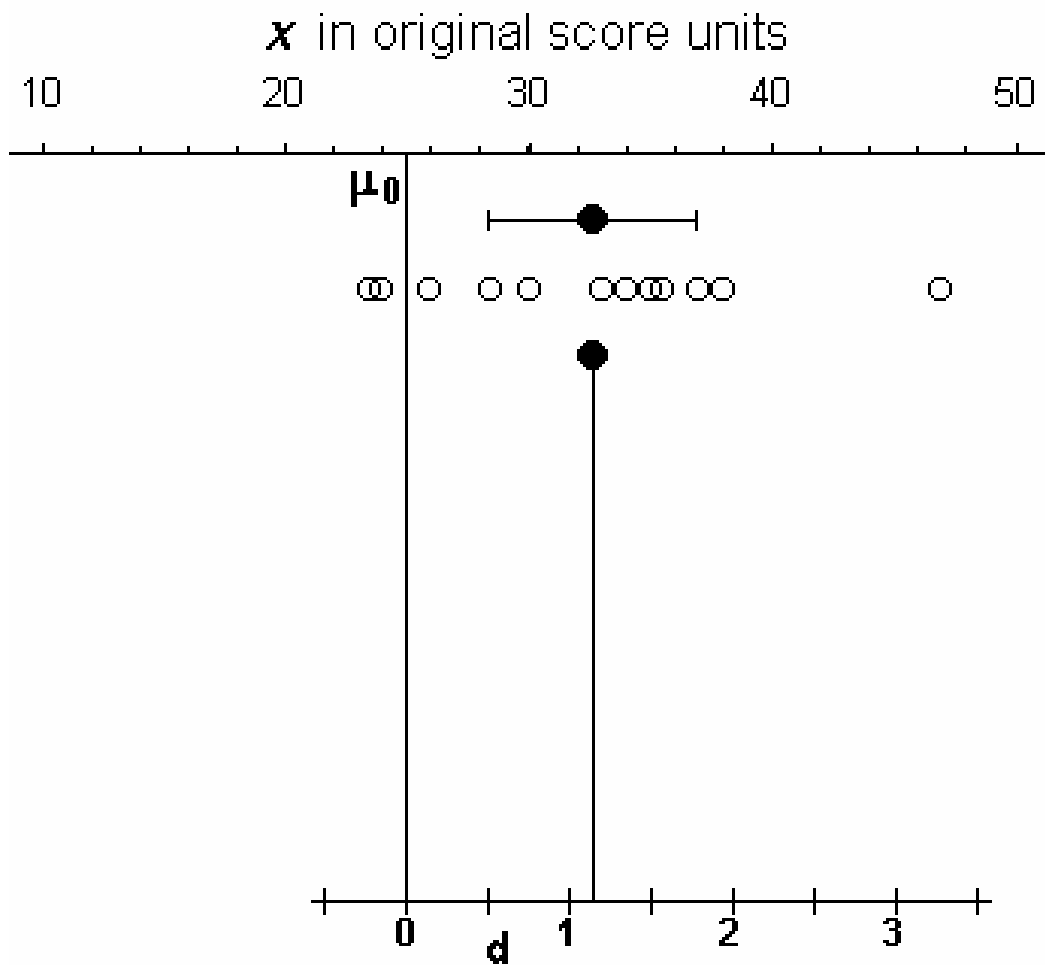


Figure 4. A part image from **CIdelta**<sup>1</sup>. To the features of Figure 1 are added a scale for  $d$ , with zero at the chosen reference value  $\mu_0$ , and which is marked in units of  $s = 6.72$ , the sample standard deviation. The lower filled dot is the mean  $d = 1.14$  for this sample.

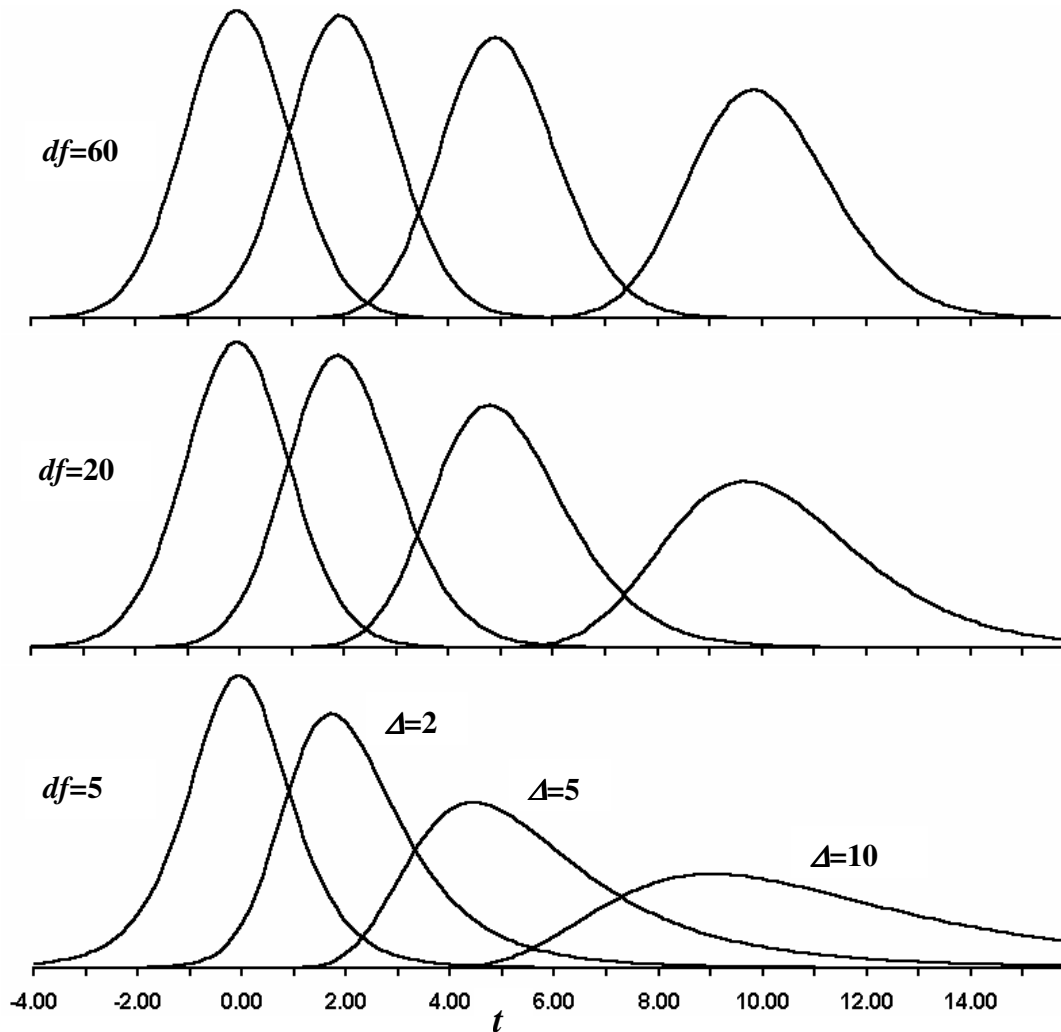


Figure 5. Central and noncentral  $t$  distributions. The  $t$  axis at the bottom is the same for all sets of curves. On each line the four curves have the same  $df$ , as indicated at left. On each line the left curve is central  $t$  ( $\Delta = 0$ ), while curves to the right have successively greater noncentrality parameters:  $\Delta = 2, 5, 10$  respectively. The curves illustrate that noncentral  $t$  distributions are centred approximately at  $\Delta$ , especially for  $\Delta$  small and  $df$  large. The variance of noncentral  $t$ , and the degree of positive skew (for  $\Delta$  positive, as here) are both larger for large  $\Delta$ , and small  $df$ . Noncentral  $t$  approaches its normal distribution asymptote only slowly. Use **NonCentralt**<sup>1</sup> to explore these properties further.

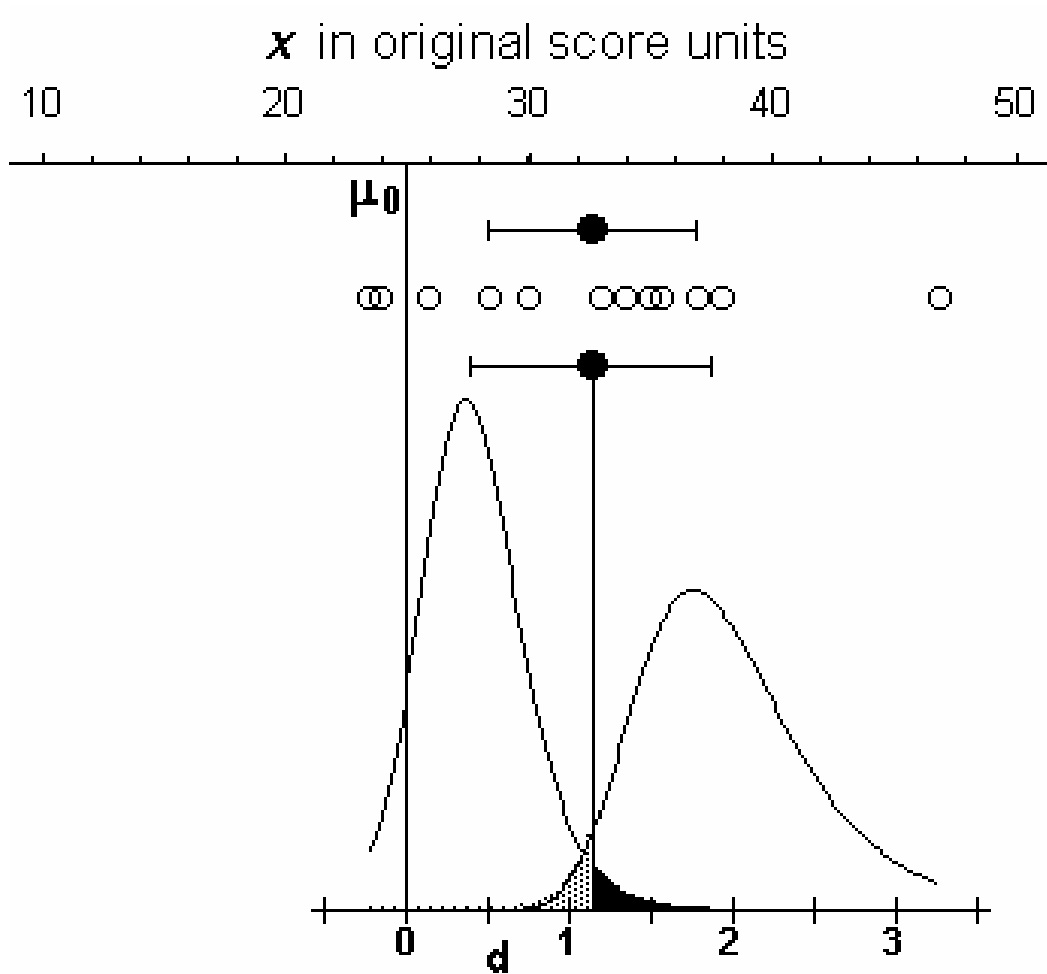


Figure 6. A part image from **CIdelta**<sup>1</sup>. To the features of Figure 4 are added two noncentral  $t$  distributions, positioned in accord with Method 2 (see text) and as illustrated (for central  $t$  distributions) by Figure 2. The  $\Delta$  values for the two curves shown (1.35, 6.43) are converted to  $\delta$  values, which become the ends of the CI for  $\delta$  (0.39, 1.86). This interval is shown as the bars around the lower filled dot, which as in Figure 4 is  $d$  for this sample. Note that the lower CI (for  $\delta$ ) refers to the lower scale, and upper CI (for  $\mu$ ) refers to the upper scale. They are different intervals, estimating different parameters.

### Confidence Intervals and Meta-analytic thinking

Assumes complete pooling

%conf 95

Study no.	$d$	$n$	$t$	$p_2$	CI for $\delta$			
1	1.5	16	6.00	0.000	***	0.76	2.21	Set CI
2	-0.2	29	-1.08	0.291	ns	-0.56	0.17	Set CI
3	0.7	36	4.20	0.000	***	0.33	1.06	Set CI
4	0.16	24	0.78	0.441	ns	-0.24	0.56	Set CI
5	0.35	40	2.21	0.033	*	0.03	0.67	Set CI
6	-0.1	20	-0.45	0.660	ns	-0.54	0.34	Set CI
7	0.3	8	0.85	0.424	ns	-0.41	1.00	Set CI
8								Set CI
9								Set CI
10								Set CI

#### Past research, pooled

0.356 173 4.69 0.000 \*\*\* 0.20 0.51 Set CI

#### Current study

0.34 25 1.70 0.102 ns -0.06 0.74 Set CI

#### Past + current, pooled

0.354 198 4.98 0.000 \*\*\* 0.21 0.50 Set CI

Set all CIs

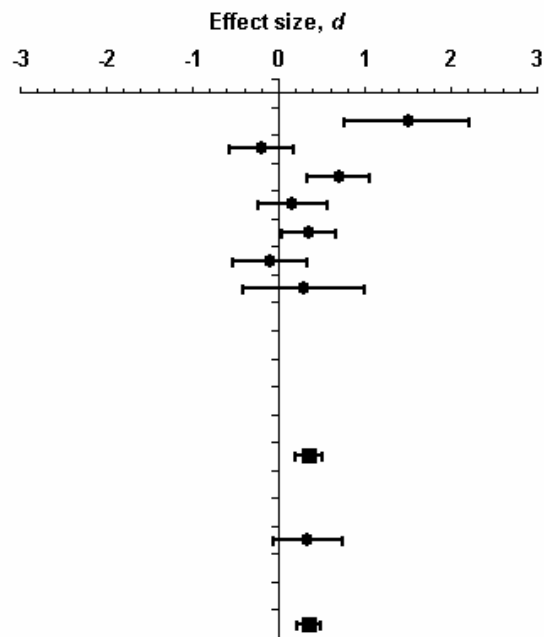
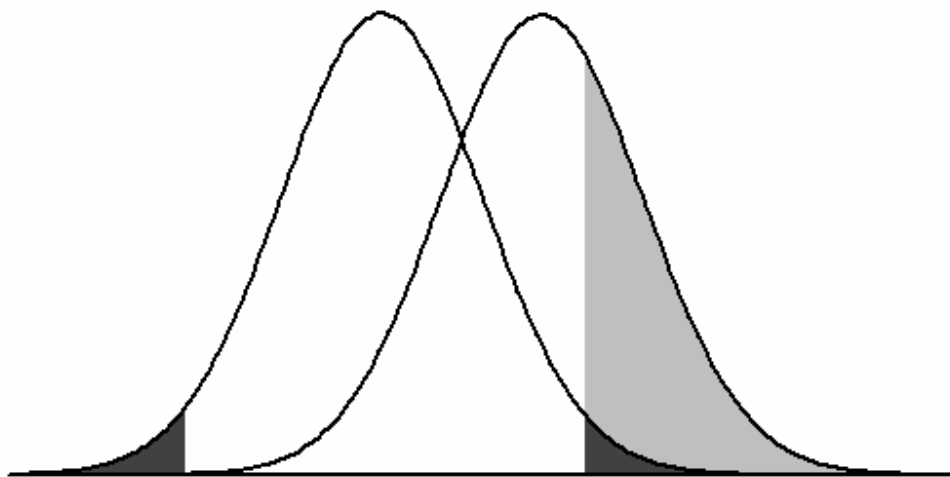
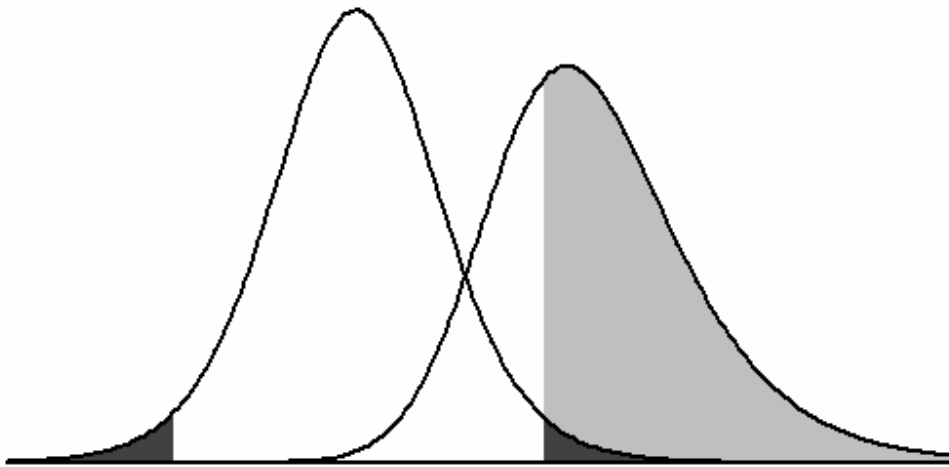


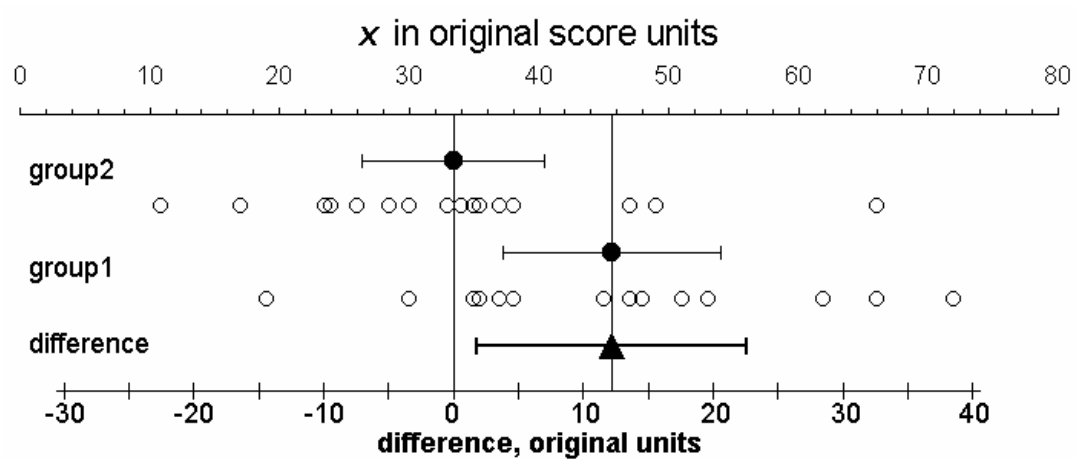
Figure 7. A part image from the standardised units sheet of **MAtHinking**<sup>1</sup>. The  $d$  and  $n$  values for 7 supposed previous studies are entered at top left. The program calculates the results of a statistical significance test for each study and shows these in the next three columns. Clicking the grey buttons triggers the iterative calculation of the CI for  $\delta$ , the results of which are shown in two columns and pictured at right as bars around the  $d$  value. A similar analysis is shown for a simple pooling of the 7 previous studies. When  $d$  and  $n$  for our supposed current study are entered, the analysis is shown for this study, and for all 8 studies pooled.



*Figure 8.* The conventional diagram for statistical power, as reproduced in many textbooks. The two sampling distribution curves appear identical, the left curve applying if  $H_0$  is true and the right if  $H_a$  is true. The dark shaded tails of the left curve correspond to the rejection region (total area  $\alpha$ , which is 0.05 in this case) for the statistical significance test. The total area of the grey tails of the right curve is statistical power. In this case, power = 0.36. Note that the left tail of the right curve makes a contribution to power although here, as in most practical situations, this is tiny. If  $\sigma$  is known and we are using a  $z$  test, the two curves are normal distributions and are indeed identical in shape. In any other case, however, the  $H_0$  curve is a  $t$  distribution while the  $H_a$  curve is a *noncentral t* distribution, and thus is different in shape from the left curve.



*Figure 9.* A part image from **Power**<sup>1</sup>, for a single sample for which  $n = 12$  and so  $df = 11$ . The left curve applies when  $H_0$  is true, and is a central  $t$  distribution. As in Figure 8 the dark tails correspond to the rejection region and have total area  $\alpha = 0.05$ . The right curve applies when  $H_a$  is true, and is a noncentral  $t$  distribution, with  $\Delta = 2.7$ , and so  $\delta = 0.78$ . The grey tails of this curve give power, which here is 0.69. Note that the two curves differ considerably in shape, and the  $H_a$  distribution is skewed, with an outward fat tail.



*Figure 10.* A part image from the Case 2 sheet of **CIoriginal**<sup>1</sup>, for two independent groups. In this case  $n_1 = 14$  and  $n_2 = 16$ . For each group the dot plot (open circles), mean (filled circle) and 95% CI for  $\mu_1$  or  $\mu_2$  is shown. The lower axis is in original units but with zero aligned with  $\bar{x}_2$ , the group 2 mean, so that it indicates  $(\bar{x}_1 - \bar{x}_2)$ , the difference between the means. The filled triangle is this difference, referred to the lower axis, and its heavy bars are the CI for  $(\mu_1 - \mu_2)$ , the difference between the population means. Note that the CI for the difference is as usual longer than the CI for either single group.