

What is an 'Effect Size'?

A guide for users

January 2000

Robert Coe

'Effect Size' is simply a way of quantifying the effectiveness of a particular intervention, relative to some comparison. It is easy to calculate, readily understood and can be applied to any measured outcome in Education or Social Science. It allows us to move beyond the simplistic, 'Does it work or not?' to the far more sophisticated, 'How well does it work in a range of contexts?' Moreover, by placing the emphasis on the most important aspect of an intervention – the size of the effect – rather than its statistical significance (which conflates Effect Size and sample size), it promotes a more scientific approach to the accumulation of knowledge. For these reasons, Effect Size is an important tool in reporting and interpreting effectiveness.

The routine use of Effect Sizes, however, has often been limited to meta-analysis – for combining and comparing estimates from different studies – and is all too rare in original reports of educational research. Formulae for the calculation of Effect Sizes do not appear in most statistics text books (other than those devoted to meta-analysis), and are seldom taught in standard research methods courses. For these reasons, even the researcher who is convinced by the wisdom of using measures of Effect Size, and is not afraid to confront the orthodoxy of conventional practice, may find that it is quite hard to know exactly how to do so.

The following guide is written for non-statisticians, though inevitably some equations and technical language have been used. It describes what Effect Size is, what it means, how it can be used and some potential problems associated with using it.

1. Why do we need 'Effect Size'?

Consider an experiment conducted by Val Dowson to investigate time of day effects on learning: do children learn better in the morning or afternoon? (ref) A group of 38 children were included in the experiment. Half were randomly allocated to listen to a story and answer questions about it at 9am, the other half to hear exactly the same story (on tape) and answer the same questions at 3pm. Their comprehension was measured by the number of questions answered correctly out of 20.

The average score was 15.2 for the morning group, 17.9 for the afternoon group: a difference of 2.7. But how big a difference is this? If the outcome were measured on a familiar scale, such as GCSE grades, interpreting the difference would not be a problem. If the average difference were, say, half a grade, most people would have a fair idea of the educational significance of the effect of reading a story at different times of day. However, in many experiments there is no familiar scale available on which to record the outcomes. The experimenter often has to invent a scale or to use (or adapt) an already existing one – but generally not one whose interpretation will be familiar to most people.

One way to get over this problem is to use the amount of variation in scores to contextualise the difference. If there were no overlap at all and every single person in the afternoon group had done better on the test than everyone in the morning group, then this would seem like a very substantial difference. On the other hand, if the spread of scores were large and the overlap much bigger than the difference between the groups, then the effect might seem less significant. Because we have an idea of the amount of variation found within a group, we can use this as a yardstick against which to compare the difference. This idea is quantified in the calculation of the *effect size*. The concept is illustrated in Figure 1, which shows two possible ways the difference might vary in relation to the overlap. If the difference were as in graph (a) it would be very significant; in graph (b), on the other hand, the difference might hardly be noticeable.

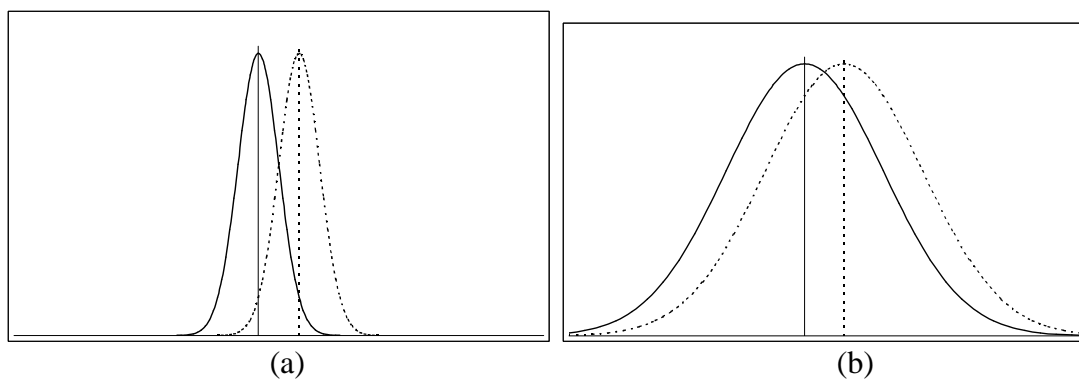


Figure 1

2. How is it calculated?

The effect size is just the standardised mean difference between the two groups. In other words:

$$\text{Effect Size} = \frac{[\text{Mean of experimental group}] - [\text{Mean of control group}]}{\text{Standard Deviation}}$$

Equation 1

If it is not obvious which of two groups is the ‘experimental’ (i.e. the one which was given the ‘new’ treatment being tested) and which the ‘control’ (the one given the ‘standard’ treatment – or no treatment – for comparison), the difference can still be calculated. In this case, the ‘effect size’ simply measures the difference between them, so it is important in quoting the effect size to say which way round the calculation was done.

The ‘standard deviation’ is a measure of the spread of a set of values.¹ Here it can refer either to the standard deviation of the control group, or to a ‘pooled’ value from both groups (see question 4, below, for more explanation of this).

In Dowson’s time of-day effects experiment, the standard deviation (SD) = 3.3, so the effect size was $(17.9 - 15.2)/3.3 = 0.8$.

3. How can effect sizes be interpreted?

One feature of an effect size is that it can be directly converted into statements about the overlap between the two samples in terms of a comparison of percentiles.

An effect size is exactly equivalent to a ‘Z-score’ of a standard Normal distribution. For example, an effect size of 0.8 means that the score of the average person in the experimental group exceeds the scores of 79% of the control group. With the two groups of 19 in the time-of-day effects experiment, the average person in the ‘afternoon’ group (i.e the one who would have been ranked 10th in the group) would have scored about the same as the 4th highest person in the ‘morning’ group. Visualising these two individuals can give quite a graphic interpretation of the difference between the two effects.

Table 1 shows conversions of effect sizes to percentiles (I_1) and the equivalent change in rank order for a group of 25 (I_2). For example, for an effect-size of 0.6, the value of 73% indicates that the average person in the experimental group would score higher than 73% of a control group that was initially equivalent. If the group consisted of 25 people, this is the same as saying that the average person (ie ranked 13th in the group) would now be on a par with the person ranked 7th in the control group. Notice that an effect-size of 1.6 would raise the average person to be level with the top ranked individual in the control group, so effect sizes larger than this are illustrated in terms of the top person in a larger group. For example, an effect size of 3.0 would bring the average person in a group of 740 level with the previously top person in the group.

Another way to conceptualise the overlap is in terms of the probability that one could guess which group a person came from, based only on their test score – or whatever value was being compared. If the effect size were 0 (ie the two groups were the same) then the probability of a correct guess would be exactly a half – or 0.50. With a difference between the two groups equivalent to an effect size of 0.3, there is still plenty of overlap, and the probability of correctly identifying the groups rises only slightly to 0.56. With an effect size of 1, the probability is now 0.69, just over a two thirds chance. These probabilities are shown in the fourth column (I_3) of Table 1. It is clear that the overlap between experimental and control groups is substantial (and therefore the probability is still close to 0.5), even when the effect-size is quite large.

A final conversion of Effect Sizes to measures of overlap is provided by Rosenthal and Rubin (1982). They suggest taking an arbitrary threshold of ‘success’ (for example the overall median score of both groups combined) and calculating the percentages of each group above this value. The difference between these two percentages is an indicator of the size of the effect. These values are shown as (I_4).

Table 1: Interpretations of effect sizes

Effect Size	Percentage of control group who would be below average person in experimental group	Rank of person in a control group of 25 who would be equivalent to the average person in experimental group	Probability that you could guess which group a person was in from knowledge of their ‘score’.	Difference in percentage ‘successful’ in each of the two groups
d	I_1	I_2	I_3	I_4
0.0	50%	13	0.50	

0.1	54%	12	0.52	
0.2	58%	11	0.54	
0.3	62%	10	0.56	
0.4	66%	9	0.58	
0.5	69%	8	0.60	
0.6	73%	7	0.62	
0.7	76%	6	0.64	
0.8	79%	6	0.66	
0.9	82%	5	0.67	
1.0	84%	4	0.69	
1.2	88%	3	0.73	
1.4	92%	2	0.76	
1.6	95%	1	0.79	
1.8	96%	1	0.82	
2.0	98%	1 (or 1 st out of 44)	0.84	
2.5	99%	1 (or 1 st out of 160)	0.89	
3.0	99.9%	1 (or 1 st out of 740)	0.93	

It should be noted that the values in Table 1 depend on the assumption of a Normal distribution. The interpretation of effect sizes in terms of percentiles is very sensitive to violations of this assumption (see below).

Another way to interpret effect sizes is to compare them to the effect sizes of differences that are familiar. For example, Cohen (1969, p23) describes an effect size of 0.2 as ‘small’ and gives to illustrate it the example that the difference between the heights of 15 year old and 16 year old girls in the US corresponds to an effect of this size. An effect size of 0.5 is described as ‘medium’ and is ‘large enough to be visible to the naked eye’. A 0.5 effect size corresponds to the difference between the heights of 14 year old and 18 year old girls. Cohen describes an effect size of 0.8 as ‘grossly perceptible and therefore large’ and equates it to the difference between the heights of 13 year old and 18 year old girls. As a further example he states that the difference in IQ between holders of the Ph.D. degree and ‘typical college freshmen’ is comparable to an effect size of 0.8.

Cohen does acknowledge the danger of using terms like ‘small’, ‘medium’ and ‘large’ out of context. Glass et al (1981, p104) are particularly critical of this approach, arguing that the effectiveness of a particular intervention can only be interpreted in relation to other interventions that seek to produce the same effect. They also point out that the practical importance of an effect depends entirely on its relative costs and benefits. In education, if it could be shown that making a small and inexpensive change would raise academic achievement by an effect size of even as little as 0.1, then this could be a very significant improvement, particularly if the improvement applied uniformly to all students, and even more so if the effect were cumulative over time.

Glass et al (1981, p102) give the example that an effect size of 1 corresponds to the difference of about a year of schooling on the performance in achievement tests of pupils in elementary (ie primary) schools. However, an analysis of a standard spelling test used in Britain (Vincent and Crumpler, 1997) suggests that the increase

in a spelling age from 11 to 12 corresponds to an effect size of about 0.3, but seems to vary according to the particular test used.

The distribution of GCSE grades in compulsory subjects (ie Maths and English) have standard deviations of between 1.5 – 1.8 grades, so an improvement of one GCSE grade represents an effect size of 0.5 – 0.7. In the context of secondary school therefore, introducing a change in practice whose effect size was known to be 0.6 would be likely to result in an improvement of about a GCSE grade for each pupil in each subject. For a school in which 50% of pupils were previously gaining five or more A* – C grades, this percentage (other things being equal, and assuming that the effect applied equally across the whole curriculum) would rise to 73%.² Even Cohen's 'small' effect of 0.2 would produce an increase from 50% to 68% – a difference that most schools would probably categorise as quite substantial.

Finally, the interpretation of effect sizes can be greatly helped by a few examples from existing research. Table 2 lists a selection of these, many of which are taken from Lipsey and Wilson (1993). The examples cited are given for illustration of the use of effect size measures; they are not intended to be the definitive judgement on the relative efficacy of different interventions. In interpreting them, therefore, one should bear in mind that most of the meta-analyses from which they are derived can be (and often have been) criticised for a variety of weaknesses, that the range of circumstances in which the effects have been found may be limited, and that the Effect Size quoted is an average which is often based on quite widely differing values.

Table 2: Examples of effect sizes from research

Intervention	Outcome	Effect Size	Source
Reducing class size from 23 to 15	Students' test performance in reading	0.30	Finn and Achilles, (1990)
	Students' test performance in maths	0.32	
Small (<30) vs large class size	Attitudes of students	0.47	Smith and Glass (1980)
	Attitudes of teachers	1.03	
Setting students vs mixed ability grouping	Student achievement (overall)	0.00	Mosteller, Light and Sachs (1996)
	Student achievement (for high-achievers)	0.08	
	Student achievement (for low-achievers)	-0.06	
Open ('child-centred') vs traditional classroom organisation	Student achievement	-0.06	Giaconia and Hedges (1982)
	Student attitudes to school	0.17	
Mainstreaming vs special education (for primary age, disabled students)	Achievement	0.44	Wang and Baker (1986)
Practice test taking	Test scores	0.32	Kulik, Bangert and Kulik (1984)
Inquiry-based vs traditional science curriculum	Achievement	0.30	Shymansky, Hedges and Woodworth (1990)
Therapy for test-anxiety (for anxious students)	Test performance	0.42	Hembree (1988)
Feedback to teachers about student performance (students with IEPs)	Student achievement	0.70	Fuchs and Fuchs (1986)
Peer tutoring	Achievement of tutees	0.40	Cohen, Kulik and Kulik, (1982)
	Achievement of tutors	0.33	
Individualised instruction	Achievement	0.10	Bangert, Kulik and Kulik (1983)
Computer assisted instruction (CAI)	Achievement (all studies)	0.24	Fletcher-Flinn and Gravatt (1995)
	Achievement (in well controlled studies)	0.02	
Additive-free diet	Children's hyperactivity	0.02	Kavale and Forness (1983)
Relaxation training	Medical symptoms	0.52	Hyman <i>et al</i> (1989)
Targeted interventions for at-risk students	Achievement	0.63	Slavin and Madden (1989)
School-based substance abuse education	Substance use	0.12	Bangert-Drowns (1988)
Treatment programmes for juvenile delinquents	Delinquency	0.17	Lipsey (1992)

4. Is it really as simple as that?

Equation 1 (above) provides the basic definition of ‘effect size’ which is suitable for most purposes. However, there are a few complications that will sometimes need to be taken into account.

Control group or pooled standard deviation?

The first problem is the issue of which ‘standard deviation’ to use. Ideally, the control group will provide the best estimate of standard deviation, since it consists of a representative group of the population who have not been affected by the experimental intervention. However, unless the control group is very large, the estimate of the ‘true’ population standard deviation derived from only the control group is likely to be appreciably less accurate than an estimate derived from both the control and experimental groups. Moreover, in studies where there is not a true ‘control’ group (for example the time-of-day effects experiment) then it may be an arbitrary decision which group’s standard deviation to use, and it will often make an appreciable difference to the estimate of effect size.

For these reasons, it is often better to use a ‘pooled’ estimate of standard deviation. The pooled estimate is essentially an average of the standard deviations of the experimental and control groups.³ Equation 2 gives the formula for its calculation:

$$SD_{\text{pooled}} = \sqrt{\frac{(N_E - 1)SD_E^2 + (N_C - 1)SD_C^2}{N_E + N_C - 2}}$$

Equation 2

(Where N_E and N_C are the numbers in the experimental and control groups, respectively, and SD_E and SD_C are their standard deviations.)

The use of a pooled estimate of standard deviation depends on the assumption that the two calculated standard deviations are estimates of *the same* population value. In other words, that the experimental and control group standard deviations differ only as a result of sampling variation. Where this assumption cannot be made (either because there is some reason to believe that the two standard deviations are likely to be systematically different, or if the actual measured values are very different), then a pooled estimate should not be used.

In the example of Dowson’s time of day experiment, the standard deviations for the morning and afternoon groups were 4.12 and 2.10 respectively. With $N_E = N_C = 19$, Equation 2 therefore gives SD_{pooled} as 3.3, which was the value used in Equation 1 to give an effect size of 0.8 (see above, page 2). However, the difference between the two standard deviations seems quite large in this case. Given that the afternoon group mean was 17.9 out of 20, it seems likely that its standard deviation may have been reduced by a ‘ceiling effect’ – ie the spread of scores was limited by the maximum available mark of 20. In this case therefore, it might be more appropriate to use the morning group’s standard deviation as the best estimate. Doing this will reduce the effect size to 0.7, and it then becomes a somewhat arbitrary decision which

value of the effect size to use. A general rule of thumb in statistics when two valid methods give different answers is: ‘If in doubt, cite both.’

Corrections for bias

Whichever version of the standard deviation is used, it can usually only be calculated from the (sometimes quite small) sample of values available. It should therefore be regarded as only an estimate of the ‘true’ population value, and subject to sampling error. Although using the pooled standard deviation to calculate the effect size generally gives a better estimate than the control group SD,⁴ it is still unfortunately slightly biased. In fact, it can be shown that when effect sizes are calculated from samples taken from a known population, on average the sample values are a little larger than the value for the population from which they come. Ideally, therefore, we should correct for this bias, particularly if the sample is small.

Hedges and Olkin⁵ give a formula which provides an approximate correction:

$$\text{Unbiased estimate of } d \cong \text{Calculated value of } d \times \left(1 - \frac{3}{\{4(N_E + N_C) - 9\}} \right)$$

Equation 3

In Dowson’s experiment with 38 values, the correction factor will be 0.98, so it makes very little difference, reducing the effect size estimate from 0.82 to 0.80. Given the likely accuracy of the figures on which this is based, it is probably only worth quoting one decimal place, so the figure of 0.8 stands.

5. What is the relationship between ‘effect size’ and ‘significance’?

Effect size quantifies the size of the difference between two groups, and may therefore be said to be a true measure of the significance of the difference. If, for example, the results of Dowson’s ‘time of day effects’ experiment were found to apply generally, we might ask the question: ‘How much difference would it make to children’s learning if they were taught a particular topic in the afternoon instead of the morning?’ The best answer we could give to this would be in terms of the effect size.

However, in statistics the word ‘significance’ is often used to mean ‘statistical significance’, which is the likelihood that the difference between the two groups could just be an accident of sampling. If you take two samples from the same population there will always be a difference between them. The statistical significance is usually calculated as a ‘p-value’, the probability that a difference of at least the same size would have arisen by chance, even if there really were no difference between the two populations. For differences between the means of two groups, this p-value would normally be calculated from a statistical procedure known as a ‘t-test’. By convention, if $p < 0.05$ (ie below 5%), the difference is taken to be large enough to be ‘significant’; if not, then it is ‘not significant’.

There are a number of problems with using ‘significance tests’ in this way. The main one is that the p-value depends essentially on two things: the size of the effect *and* the size of the sample. One would get a ‘significant’ result either if the effect were very big (despite having only a small sample) or if the sample were very big (even if the actual effect size were tiny). It is important to know the statistical significance of a result, since without it there is a danger of drawing firm conclusions

from studies where the sample is too small to justify such confidence. However, statistical significance does *not* tell you the most important thing: *the size of the effect*. One way to overcome this confusion is to report the effect size, together with an estimate of its likely ‘margin for error’ or ‘confidence interval’.

6. What is the margin for error in estimating effect sizes?

Clearly, if an effect size is calculated from a very large sample it is likely to be more accurate than one calculated from a small sample. This ‘margin for error’ can be quantified using the idea of a ‘confidence interval’, which provides the same information as is usually contained in a significance test: using a ‘95% confidence interval’ is equivalent to taking a ‘5% significance level’. To calculate a 95% confidence interval, you assume that the value you got (e.g. the effect size estimate of 0.8) is the ‘true’ value, but calculate the amount of variation in this estimate you would get if you repeatedly took new samples of the same size (i.e. different samples of 38 children). For every 100 of these hypothetical new samples, 95 would give estimates of the effect size within the ‘95% confidence interval’. If this confidence interval includes zero, then that is the same as saying that the result is not statistically significant. If, on the other hand, zero is outside the range, then it is ‘statistically significant at the 5% level’. Using a confidence interval is a better way of conveying this information since it keeps the emphasis on the effect size – which is the important information – rather than the p-value.

A formula for calculating the confidence interval for an effect size is given by Hedges and Olkin⁶. If the effect size estimate from the sample is d , then it is Normally distributed, with standard deviation:

$$\sigma[d] = \sqrt{\frac{N_E + N_C}{N_E \times N_C} + \frac{d^2}{2(N_E + N_C)}}$$

Equation 4

(Where N_E and N_C are the numbers in the experimental and control groups, respectively.)

Hence a 95% confidence interval⁷ for d would be from

$$d - 1.96 \times \sigma[d] \quad \text{to} \quad d + 1.96 \times \sigma[d]$$

Equation 5

To use the figures from the time-of-day experiment again, $N_E = N_C = 19$ and $d = 0.8$, so $\sigma[d] = \sqrt{(0.105 + 0.008)} = 0.34$. Hence the 95% confidence interval is [0.14, 1.46]. This would normally be interpreted⁸ as meaning that the ‘true’ effect of time-of-day is very likely to be between 0.14 and 1.46. In other words, it is almost certainly positive (ie afternoon is better than morning) and the difference may well be quite large.

7. How can knowledge about effect sizes be combined?

One of the main advantages of using effect size is that when a particular experiment has been replicated, the different effect size estimates from each study can easily be combined to give an overall best estimate of the size of the effect. This process of synthesising experimental results into a single effect size estimate is known as ‘meta-analysis’. It was developed by an educational statistician, Gene Glass,⁹ and is now widely used, not only in education, but in medicine and throughout the social sciences. A brief and accessible introduction to the idea of meta-analysis can be found in Fitz-Gibbon (1984).

Meta-analysis, however, can do much more than simply produce an overall ‘average’ effect size, important though this often is. If, for a particular intervention, some studies produced large effects, and some small effects, it would be of limited value simply to combine them together and say that the average effect was ‘medium’. Much more useful would be to examine the original studies for any differences between those with large and small effects and to try to understand what factors might account for the difference. The best meta-analysis, therefore, involves seeking relationships between effect sizes and characteristics of the intervention, the context and study design in which they were found.¹⁰

The importance of replication in gaining evidence about what works cannot be overstressed. In Dowson’s time-of-day experiment the effect was found to be large enough to be statistically and educationally significant. Because we know that the pupils were allocated randomly to each group, we can be confident that chance initial differences between the two groups are very unlikely to account for the difference in the outcomes. Furthermore, the use of a pre-test of both groups before the intervention makes this even less likely. However, we cannot rule out the possibility that the difference arose from some characteristic peculiar to the children in this particular experiment. For example, if none of them had had any breakfast that day, this might account for the poor performance of the morning group. However, the result would then presumably not generalise to the wider population of school students, most of whom would have had some breakfast. Alternatively, the effect might depend on the age of the students. Dowson’s students were aged 7 or 8; it is quite possible that the effect could be diminished or reversed with older (or younger) students. This illustrates the danger of implementing policy on the basis of a single experiment. Confidence in the generality of a result can only follow widespread replication.

An important consequence of the capacity of meta-analysis to combine results is that even small studies can make a significant contribution to knowledge. The kind of experiment that can be done by a single teacher in a school might involve a total of fewer than 30 students. Unless the effect is huge, a study of this size is most unlikely to get a statistically significant result. According to conventional statistical wisdom, therefore, the experiment is not worth doing. However, if the results of several such experiments are combined using meta-analysis, the overall result is likely to be highly statistically significant. Moreover, it will have the important strengths of being derived from a range of contexts (thus increasing confidence in its generality) and from real-life working practice (thereby making it more likely that the policy is feasible and can be implemented authentically).

One final caveat should be made here about the danger of combining incommensurable results. Given two (or more) numbers, one can always calculate an average. However, if they are effect-sizes from experiments that differ significantly in terms of the outcome measures used, then the result may be totally meaningless. It

can be very tempting, once effect-sizes have been calculated, to treat them as all the same and lose sight of their origins. Certainly, there are plenty of examples of meta-analyses in which the juxtaposition of effect-sizes is somewhat questionable.

8. What other factors can influence effect size?

Although effect size is a simple and readily interpreted measure of effectiveness, it can also be sensitive to a number of spurious influences, so some care needs to be taken in its use.

Restricted range

Suppose the time-of-day effects experiment were to be repeated, once with the top set in a grammar school and again with a mixed-ability group in a comprehensive. If students were allocated to morning and afternoon groups at random, the respective differences between them might be the same in each case; both means in the grammar school might be higher, but the difference between the two groups could be the same as the difference in the comprehensive. However, it is unlikely that the standard deviations would be the same. The top set of a grammar school is a highly selected group, and the spread of scores found within it would be much less than that in a true cross-section of the population, as for example in a mixed-ability comprehensive class. This, of course, would have a substantial impact on the calculation of the effect size. With the highly restricted range found in the Grammar school, the effect size would be much larger than that found in the Comprehensive.

Ideally, in calculating effect-size one should use the standard deviation of the full population, in order to make comparisons fair. However, there will be many cases in which unrestricted values are not available, either in practice or in principle. For example, in considering the effect of an intervention with university students, or with pupils with reading difficulties, one must remember that these are restricted populations. In reporting the effect-size, one should draw attention to this fact; if the amount of restriction can be quantified it may be possible to make allowance for it. Any comparison with effect sizes calculated from a full-range population must be made with great caution, if at all.

Non-Normal distributions

The interpretations of effect-sizes given in Table 1 (page 3) depend on the assumption that both control and experimental groups have a 'Normal' distribution, ie the familiar 'bell-shaped' curve, shown, for example, in Figure 1. Needless to say, if this assumption is not true then the interpretation may be altered, and in particular, it may be difficult to make a fair comparison between an effect-size based on Normal distributions and one based on non-Normal distributions.

An illustration of this is given in Figure 2, which shows the frequency curves for two distributions, one of them Normal, the other similar in shape over the central part of its range, but with somewhat fatter extremes. In fact, the latter does look just a little more spread-out than the Normal distribution, but its standard deviation is actually fifty percent higher. The consequence of this in terms of effect-size differences is shown in Figure 3. Both graphs show distributions that differ by an effect-size equal to 1, but the appearance of the effect-size difference from the graphs is rather dissimilar. In graph (b), the separation between experimental and control groups seems much larger, yet the effect-size is actually the same as for the Normal distributions plotted in graph (a). In terms of the amount of overlap, in graph (b) 94%

of the 'experimental' group are above the control group mean, compared with the value of 84% for the Normal distribution of graph (a) (as given in Table 1). This is quite a substantial difference and illustrates the danger of using the values in Table 1 when the distribution is not known to be Normal.

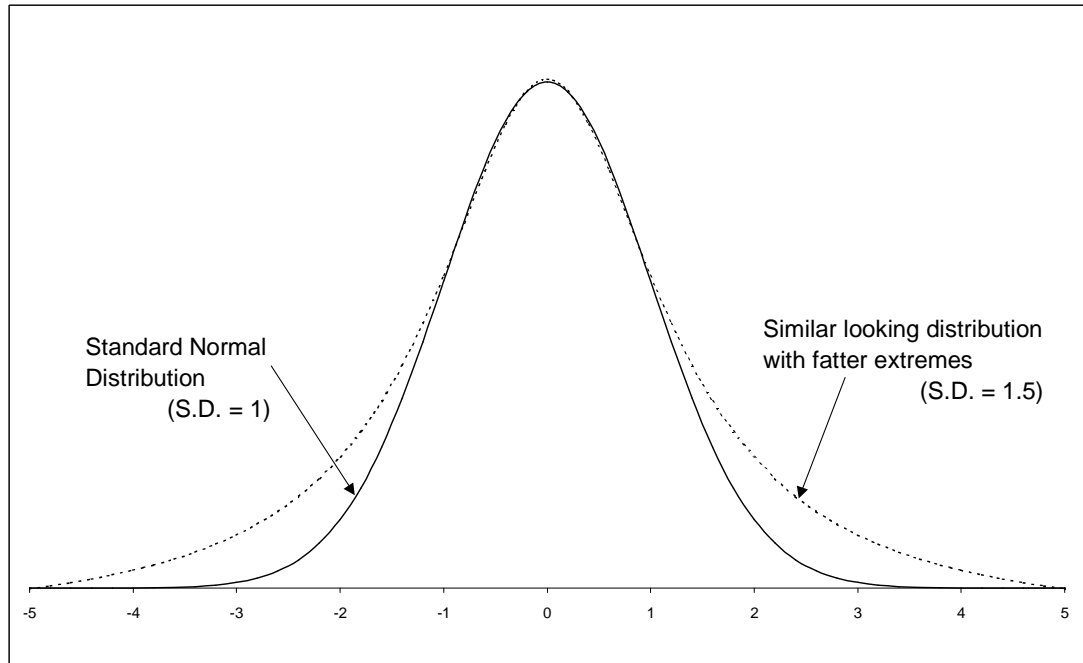


Figure 2: Comparison of Normal and non-Normal distributions

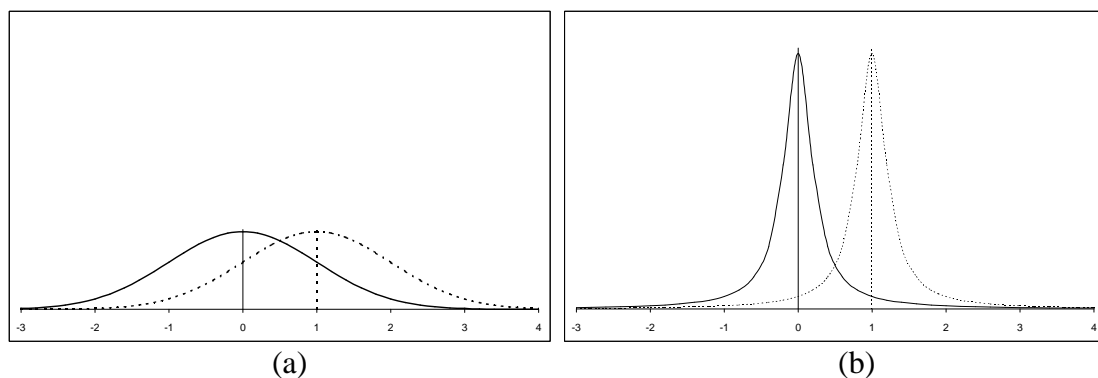


Figure 3: Normal and non-Normal distributions with effect-size = 1

Measurement reliability

A third factor that can spuriously affect an effect-size is the reliability of the measurement on which it is based. 'Reliability' here refers to a measure's accuracy, robustness and stability, often defined in terms of the extent to which two measures of the same thing (or a repeated measure) agree with each other.¹¹ To put it another way, any measure of a particular outcome may be considered to consist of the 'true' underlying value, together with a component of 'error'. The problem is that the amount of variation in measured scores for a particular sample (ie its standard

deviation) will depend on both the variation in underlying scores and the amount of error in their measurement.

To give an example, imagine the time-of-day experiment were conducted twice with two (hypothetically) identical samples of students. In the first version the test used to assess their comprehension consisted of just 10 items and their scores were converted into a percentage. In the second version a test with 50 items was used, and again converted to a percentage. The two tests were of equal difficulty and the actual effect of the difference in time-of-day was the same in each case, so the respective mean percentages of the morning and afternoon groups were the same for both versions. However, it is almost always the case that a longer test will be more reliable, and hence the standard deviation of the percentages on the 50 item test will be lower than the standard deviation for the 10 item test. Thus, although the true effect was the same, the calculated effect-sizes will be different.

In interpreting an effect-size, it is therefore important to know the reliability of the measurement from which it was calculated. This is one reason why the reliability of any outcome measure used should be reported. It is theoretically possible to make a correction for unreliability, which gives an estimate of what the effect-size would have been, had the reliability of the test been perfect. However, in practice the effect of this is rather alarming, since the worse the test was, the more you increase the estimate of the effect-size.

9. Are there alternative measures of effect-size?

Conclusions

Advice on the use of effect-sizes can be summarised as follows:

- Effect-size is a standardised, scale-free measure of the relative size of the effect of an intervention. It is particularly useful for quantifying effects measured on unfamiliar or arbitrary scales and for comparing the relative sizes of effects from different studies.
- Interpretation of effect-size generally depends on the assumptions that ‘control’ and ‘experimental’ group values are Normally distributed and have the same standard deviations. Effect-sizes can be interpreted in terms of the percentiles or ranks at which two distributions overlap, in terms of the likelihood of identifying the source of a value, or with reference to known effects or outcomes.
- Use of an effect size with a confidence interval conveys the same information as a test of statistical significance, but with the emphasis on the significance of the effect, rather than the sample size.
- Effect sizes should be calculated and reported in primary studies as well as in meta-analyses.
- Interpretation of standardised effect sizes can be problematic when a sample has restricted range or does not come from a Normal distribution, or if the measurement from which it was derived has unknown reliability.
- The use of an ‘unstandardised’ mean difference (ie the raw difference between the two groups, together with a confidence interval) may be preferable when:
 - the outcome is measured on a familiar scale

- the sample has a restricted range
 - the parent population is significantly non-Normal
 - control and experimental groups have appreciably different standard deviations
 - the outcome measure has very low or unknown reliability
- Care must be taken in comparing or aggregating effect sizes based on different measures.

¹ Various formulae for calculating standard deviation will be found in any statistics text, and are built into spreadsheets such as Excel. One simple formula for a set of values, X_1, X_2, \dots, X_n , with mean M_x is:

$$SD = \sqrt{\{ [(X_1^2 + X_2^2 + \dots + X_n^2) - n \times (M_x)^2] / (n-1) \}}$$

² This calculation is derived from a probit transformation (Glass et al, 1981, p136), based on the assumption of an underlying normally distributed variable measuring academic attainment, some threshold of which is equivalent to a student achieving 5+ A* – Cs. Percentages for the change from a starting value of 50% for other effect size values can be read directly from Table 1. Alternatively, if $\Phi(z)$ is the standard normal cumulative distribution function, p_1 is the proportion achieving a given threshold and p_2 the proportion to be expected after a change with effect size, d , then,

$$p_2 = \Phi\{\Phi^{-1}(p_1) + d\}$$

³ Note that this is not the same as the standard deviation of all the values in both groups ‘pooled’ together. If, for example each group had a low standard deviation but the two means were substantially different, the true pooled estimate (as calculated by Equation 2) would be much lower than the value obtained by pooling all the values together and calculating the standard deviation.

⁴ Both the bias and variance of the effect size estimator based on pooled SD are lower than those for control group SD, given the assumption of equal variance (Hedges and Olkin, 1985, p79).

⁵ Hedges, and Olkin, (1985) page 80

⁶ Hedges, and Olkin, (1985) page 86

⁷ The ‘1.96’ in this formula comes from the two-tailed critical value of the standard Normal Distribution. Other values may be substituted to give different confidence intervals.

⁸ In fact, this interpretation is really only justified with the incorporation of an additional ‘other things being equal’ clause and a ‘Bayesian’ approach that most statisticians explicitly rule out. However, the interpretation of precisely what significance tests do mean is controversial and often confused. See Oakes (1986) for an enlightening discussion of the issue.

⁹ See Glass, McGaw and Smith, 1981.

¹⁰ Rubin, 1992

¹¹ For further definition and discussion of the concept of reliability see, for example Kerlinger (1986) p404.

References

- Bangert, R.L., Kulik, J.A. and Kulik, C.C. (1983) ‘Individualised systems of instruction in secondary schools.’ *Review of Educational Research*, 53, 143-158.
- Bangert-Drowns, R.L. (1988) ‘The effects of school-based substance abuse education: a meta-analysis’. *Journal of Drug Education*, 18, 3, 243-65.
- Cohen, J. (1969) *Statistical Power Analysis for the Behavioral Sciences*. NY: Academic Press.

- Cohen, P.A., Kulik, J.A. and Kulik, C.C. (1982) 'Educational outcomes of tutoring: a meta-analysis of findings.' *American Educational Research Journal*, 19, 237-248.
- Finn and Achilles (1990) 'Some questions and answers about class size' *AERJ?*
- Fitz-Gibbon C.T. (1984) 'Meta-analysis: an explication'. *British Educational Research Journal*, 10, 2, 135-144.
- Fletcher-Flinn, C. and Gravatt, (1995)
- Fuchs, L.S. and Fuchs, D. (1986) 'Effects of systematic formative evaluation: a meta-analysis.' *Exceptional Children*, 53, 199-208.
- Giaconia, R.M. and Hedges, L.V. (1982) 'Identifying features of effective open education.' *Review of Educational Research*, 52, 579-602.
- Glass, G.V., McGaw, B. and Smith, M.L. (1981) *Meta-Analysis in Social Research*. London: Sage.
- Hedges, L. and Olkin, I. (1985) *Statistical Methods for Meta-Analysis*. New York: Academic Press.
- Hembree, R. (1988) 'Correlates, causes effects and treatment of test anxiety.' *Review of Educational Research*, 58(1), 47-77.
- Hyman, R.B, Feldman, H.R., Harris, R.B., Levin, R.F. and Malloy, G.B. (1989) 'The effects of relaxation training on medical symptoms: a meat-analysis.' *Nursing Research*, 38, 216-220.
- Kavale, K.A. and Forness, S.R. (1983) 'Hyperactivity and diet treatment: a meat-analysis of the Feingold hypothesis.' *Journal of Learning Disabilities*, 16, 324-330.
- Kerlinger, F.N. (1986) *Foundations of Behavioral Research*, New York: Holt, Rinehart and Winston.
- Kulik, J.A., Kulik, C.C. and Bangert, R.L. (1984) 'Effects of practice on aptitude and achievement test scores.' *American Education Research Journal*, 21, 435-447.
- Lipsey, M.W. (1992) 'Juvenile delinquency treatment: a meta-analytic inquiry into the variability of effects.' In T.D. Cook, H. Cooper, D.S. Cordray, H. Hartmann, L.V. Hedges, R.J. Light, T.A. Louis and F. Mosteller (Eds) *Meta-analysis for explanation*. New York: Russell Sage Foundation.
- Mosteller, F., Light, R.J. and Sachs (1996)
- Oakes, M. (1986) *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. New York: Wiley.
- Rosenthal, R, and Rubin, D.B. (1982) 'A simple, general purpose display of magnitude of experimental effect.' *Journal of Educational Psychology*, 74, 166-169.
- Rubin, D.B. (1992) 'Meta-analysis: literature synthesis or effect-size surface estimation.' *Journal of Educational Statistics*, 17, 4, 363-374.
- Shymansky, J.A., Hedges, L.V. and Woodworth, G. (1990) A reassessment of the effects of inquiry-based science curricula of the 60's on student performance.' *Journal of Research in Science Teaching*, 27, 127-144.

Slavin, R.E. and Madden, N.A. (1989) 'What works for students at risk? a research synthesis.' *Educational Leadership*, 46(4), 4-13.

Smith, M.L. and Glass, G.V. (1980) 'Meta-analysis of research on class size and its relationship to attitudes and instruction.' *American Educational Research Journal*, 17, 419-433.

Vincent and Crumpler (1997)

Wang, M.C. and Baker, E.T. (1986) 'Mainstreaming programs: Design features and effects.' *Journal of Special Education*, 19, 503-523.