It's the Effect Size, Stupid
What effect size is and why it is important

Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England, 12-14 September 2002

Robert Coe

School of Education, University of Durham, Leazes Road, Durham DH1 1TA
Tel 0191 374 4504; Fax 0191 374 1900; Email r.j.coe@dur.ac.uk

Abstract

Effect size is a simple way of quantifying the difference between two groups that has many advantages over the use of tests of statistical significance alone. Effect size emphasises the size of the difference rather than confounding this with sample size. However, primary reports rarely mention effect sizes and few textbooks, research methods courses or computer packages address the concept. This paper provides an explication of what an effect size is, how it is calculated and how it can be interpreted. The relationship between effect size and statistical significance is discussed and the use of confidence intervals for the latter outlined. Some advantages and dangers of using effect sizes in meta-analysis are discussed and other problems with the use of effect sizes are raised. A number of alternative measures of effect size are described. Finally, advice on the use of effect sizes is summarised.

> *During 1992 Bill Clinton and George Bush Snr. were fighting for the presidency of the United States. Clinton was barely holding on to his place in the opinion polls. Bush was pushing ahead drawing his on his stature as an experienced world leader. James Carville, one of Clinton's top advisers decided that their push for presidency needed focusing. Drawing on the research he had conducted he came up with a simple focus for their campaign. Every opportunity he had, Carville wrote four words - 'It's the economy, stupid' - on a whiteboard for Bill Clinton to see every time he went out to speak.*

'Effect size' is simply a way of quantifying the size of the difference between two groups. It is easy to calculate, readily understood and can be applied to any measured outcome in Education or Social Science. It is particularly valuable for quantifying the effectiveness of a particular intervention, relative to some comparison. 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 generally been limited to meta-analysis – for combining and comparing estimates from different studies – and is all

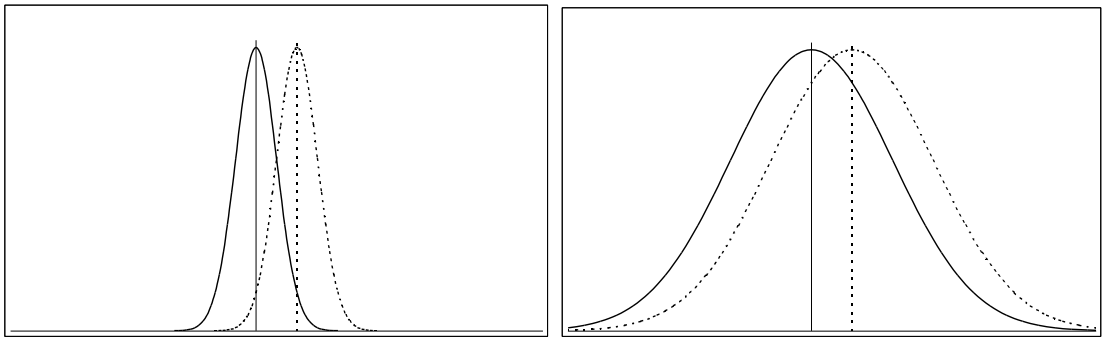
too rare in original reports of educational research (Keselman *et al.*, 1998). This is despite the fact that measures of effect size have been available for at least 60 years (Huberty, 2002), and the American Psychological Association has been officially encouraging authors to report effect sizes since 1994 – but with limited success (Wilkinson *et al.*, 1999). Formulae for the calculation of effect sizes do not appear in most statistics text books (other than those devoted to meta-analysis), are not featured in many statistics computer packages 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 Dowson (2000) to investigate time of day effects on learning: do children learn better in the morning or afternoon? A group of 38 children were included in the experiment. Half were randomly allocated to listen to a story and answer questions about it (on tape) at 9am, the other half to hear exactly the same story 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.



(a) (b)

Figure 1

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.

2. How is it calculated?

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

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

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. Here it refers to the standard deviation of the population from which the different treatment groups were taken. In practice, however, this is almost never known, so it must be estimated either from the standard deviation of the control group, or from a 'pooled' value from both groups (see question 7, below, for more discussion 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 is 0.8 standard deviations above the average person in the control group, and hence 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 I shows conversions of effect sizes (column 1) to percentiles (column 2) and the equivalent change in rank order for a group of 25 (column 3). 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 (i.e. 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.

Table I: 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'.	Equivalent correlation, *r* (=Difference in percentage 'successful' in each of the two groups, BESD)	Probability that person from experimental group will be higher than person from control, if both chosen at random (=CLES)
0.0	50%	13th	0.50	0.00	0.50
0.1	54%	12th	0.52	0.05	0.53
0.2	58%	11th	0.54	0.10	0.56
0.3	62%	10th	0.56	0.15	0.58
0.4	66%	9th	0.58	0.20	0.61
0.5	69%	8th	0.60	0.24	0.64
0.6	73%	7th	0.62	0.29	0.66
0.7	76%	6th	0.64	0.33	0.69
0.8	79%	6th	0.66	0.37	0.71
0.9	82%	5th	0.67	0.41	0.74
1.0	84%	4th	0.69	0.45	0.76
1.2	88%	3rd	0.73	0.51	0.80
1.4	92%	2nd	0.76	0.57	0.84
1.6	95%	1st	0.79	0.62	0.87
1.8	96%	1st	0.82	0.67	0.90
2.0	98%	1st (or 1st out of 44)	0.84	0.71	0.92
2.5	99%	1st (or 1st out of 160)	0.89	0.78	0.96
3.0	99.9%	1st (or 1st out of 740)	0.93	0.83	0.98

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 (i.e. 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 of Table

I. 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 slightly different way to interpret effect sizes makes use of an equivalence between the standardised mean difference (*d*) and the correlation coefficient, *r*. If group membership is coded with a dummy variable (e.g. denoting the control group by 0 and the experimental group by 1) and the correlation between this variable and the outcome measure calculated, a value of *r* can be derived. By making some additional assumptions, one can readily convert *d* into *r* in general, using the equation $r^2 = d^2 / (4+d^2)$ (see Cohen, 1969, pp20-22 for other formulae and conversion table). Rosenthal and Rubin (1982) take advantage of an interesting property of *r* to suggest a further interpretation, which they call the binomial effect size display (BESD). If the outcome measure is reduced to a simple dichotomy (for example, whether a score is above or below a particular value such as the median, which could be thought of as 'success' or 'failure'), *r* can be interpreted as the difference in the proportions in each category. For example, an effect size of 0.2 indicates a difference of 0.10 in these proportions, as would be the case if 45% of the control group and 55% of the treatment group had reached some threshold of 'success'. Note, however, that if the overall proportion 'successful' is not close to 50%, this interpretation can be somewhat misleading (Strahan, 1991; McGraw, 1991). The values for the BESD are shown in column 5.

Finally, McGraw and Wong (1992) have suggested a 'Common Language Effect Size' (CLES) statistic, which they argue is readily understood by non-statisticians (shown in column 6 of Table I). This is the probability that a score sampled at random from one distribution will be greater than a score sampled from another. They give the example of the heights of young adult males and females, which differ by an effect size of about 2, and translate this difference to a CLES of 0.92. In other words 'in 92 out of 100 blind dates among young adults, the male will be taller than the female' (p361).

It should be noted that the values in Table I 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 question 7, 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.

Table II: Examples of average 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 *et al.* (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)

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 (i.e. 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.

In England, the distribution of GCSE grades in compulsory subjects (i.e. 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 schools therefore, introducing a change in practice whose effect size was known to be 0.6 would 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%.[1] Even Cohen's 'small' effect of 0.2 would produce an increase from 50% to 58% – a difference that most schools would probably categorise as quite substantial. Olejnik and Algina (2000) give a similar example based on the Iowa Test of Basic Skills

Finally, the interpretation of effect sizes can be greatly helped by a few examples from existing research. Table II 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.

It seems to be a feature of educational interventions that very few of them have effects that would be described in Cohen's classification as anything other than 'small'. This appears particularly so for effects on student achievement. No doubt this is partly a result of the wide variation found in the population as a whole, against which the measure of effect size is calculated. One might also speculate that achievement is harder to influence than other outcomes, perhaps because most schools are already using optimal strategies, or because different strategies are likely to be effective in different situations – a complexity that is not well captured by a single average effect size.

4. 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 't-test'. By convention, if $p < 0.05$ (i.e. 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 (see, for example Cohen, 1994; Harlow *et al*., 1997; Thompson, 1999). 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'.

5. 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, by definition, 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 Olkin (1985, p86). 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 2

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

Hence a 95% confidence interval for *d* would be from

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

Equation 3

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 interpreted (despite the fact that such an interpretation is not strictly justified – see Oakes, 1986 for an enlightening discussion of this) 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 (i.e. afternoon is better than morning) and the difference may well be quite large.

6. 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 in its current form by an educational statistician, Gene Glass (See Glass *et al*., 1981) though the roots of meta-analysis can be traced a good deal further back(see Lepper *et al.*, 1999), 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 (Rubin, 1992; see also Lepper *et al.* (1999) for a discussion of the problems that can be created by failing to do this, and some other limitations of the applicability of meta-analysis).

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.

In comparing (or combining) effect sizes, one should therefore consider carefully whether they relate to the same outcomes. This advice applies not only to meta-analysis, but to any other comparison of effect sizes. Moreover, because of the sensitivity of effect size estimates to reliability and range restriction (see below), one should also consider whether those outcome measures are derived from the same (or sufficiently similar) instruments and the same (or sufficiently similar) populations.

It is also important to compare only like with like in terms of the treatments used to create the differences being measured. In the education literature, the same name is often given to interventions that are actually very different, for example, if they are operationalised differently, or if they are simply not well enough defined for it to be clear whether they are the same or not. It could also be that different studies have used the same well-defined and operationalised treatments, but the actual implementation differed, or that the same treatment may have had different levels of intensity in different studies. In any of these cases, it makes no sense to average out their effects.

7. 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. Some of these issues are outlined here.

Which '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 (Equation 4). 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 4) would be much lower than the value obtained by pooling all the values together and calculating the standard deviation. The implications of choices about which standard deviation to use are discussed by Olejnik and Algina (2000).

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

Equation 4

(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. 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' – i.e. 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

Although using the pooled standard deviation to calculate the effect size generally gives a better estimate than the control group SD, it is still unfortunately slightly biased and in general gives a value slightly larger than the true population value (Hedges and Olkin, 1985). Hedges and Olkin (1985, p80) give a formula which provides an approximate correction to this bias.

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. In fact, the correction only becomes significant for small samples, in which the accuracy is anyway much less. It is therefore hardly worth worrying about it in primary reports of empirical results. However, in meta-analysis, where results from primary studies are combined, the correction is important, since without it this bias would be accumulated.

Restricted range

Suppose the time-of-day effects experiment were to be repeated, once with the top set in a highly selective 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 selective 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 spread of scores found

within the highly selected group would be much less than that in a true cross-section of the population, as for example in the 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 selective 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 I depend on the assumption that both control and experimental groups have a 'Normal' distribution, i.e. 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.

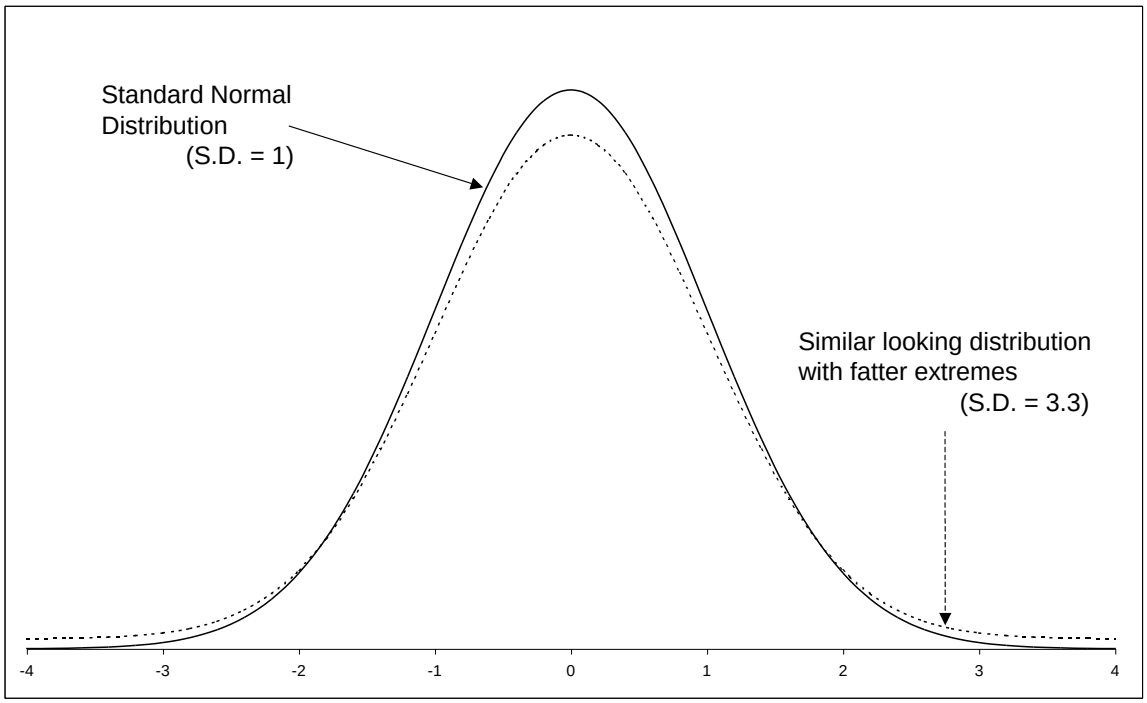
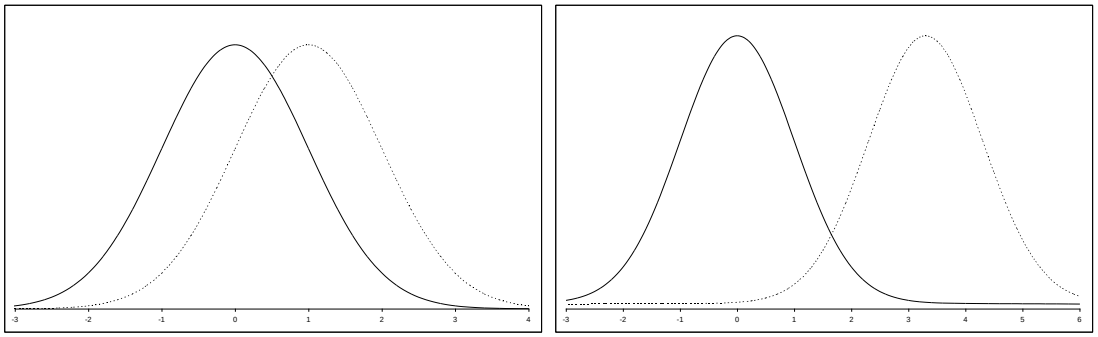


Figure 2: Comparison of Normal and non-Normal distributions

An illustration of this is given in Figure 2, which shows the frequency curves for two distributions, one of them Normal, the other a 'contaminated normal' distribution (Wilcox, 1998), which is similar in shape, 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 over three times as big. 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) 97% 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 I). This is quite a substantial difference and illustrates the danger of using the values in Table I when the distribution is not known to be Normal.



(a) (b)

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. According to classical measurement theory, 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 (i.e. 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 (sometimes called 'attenuation'), 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. Moreover, estimates of reliability are dependent on the particular population in which the test was used, and are themselves anyway subject to sampling error. For further discussion of the impact of reliability on effect sizes, see Baugh (2002).

8. Are there alternative measures of effect-size?

A number of statistics are sometimes proposed as alternative measures of effect size, other than the 'standardised mean difference'. Some of these will be considered here.

Proportion of variance accounted for

If the correlation between two variables is 'r', the square of this value (often denoted with a capital letter: R^2) represents the proportion of the variance in each that is 'accounted for' by the other. In other words, this is the proportion by which the variance of the outcome measure is reduced when it is replaced by the variance of the residuals from a regression equation. This idea can be extended to multiple regression (where it represents the proportion of the variance accounted for by all the independent variables together) and has close analogies in ANOVA (where it is usually called 'eta-squared', η^2). The calculation of *r* (and hence R^2) for the kind of experimental situation we have been considering has already been referred to above.

Because R^2 has this ready convertibility, it (or alternative measures of variance accounted for) is sometimes advocated as a universal measure of effect size (e.g. Thompson, 1999). One disadvantage of such an approach is that effect size measures based on variance accounted for suffer from a number of technical limitations, such as sensitivity to violation of assumptions (heterogeneity of variance, balanced designs) and their standard errors can be large (Olejnik and Algina, 2000). They are also generally more statistically complex and hence perhaps less easily understood. Further, they are non-directional; two studies with precisely opposite results would report exactly the same variance accounted for. However, there is a more fundamental objection to the use of what is essentially a measure of association to indicate the strength of an 'effect'.

Expressing different measures in terms of the same statistic can hide important differences between them; in fact, these different 'effect sizes' are fundamentally different, and should not be confused. The crucial difference between an effect size calculated from an experiment and one calculated from a correlation is in the causal nature of the claim that is being made for it. Moreover, the word 'effect' has an inherent implication of causality: talking about 'the effect of A on B' does suggest a causal relationship rather than just an association. Unfortunately, however, the word 'effect' is often used when no explicit causal claim is being made, but its implication is sometimes allowed to float in and out of the meaning, taking advantage of the ambiguity to suggest a subliminal causal link where none is really justified.

This kind of confusion is so widespread in education that it is recommended here that the word 'effect' (and therefore 'effect size') should not be used unless a deliberate and explicit causal claim is being made. When no such claim is being made, we may talk about the 'variance accounted for' (R^2) or the 'strength of association' (r), or simply – and perhaps most informatively – just cite the regression coefficient (Tukey, 1969). If a causal claim is being made it should be explicit and justification provided. Fitz-Gibbon (2002) has recommended an alternative approach to this problem. She has suggested a system of nomenclature for different kinds of effect sizes that clearly distinguishes between effect sizes derived from, for example, randomised-controlled, quasi-experimental and correlational studies.

Other measures of effect size

It has been shown that the interpretation of the 'standardised mean difference' measure of effect size is very sensitive to violations of the assumption of normality. For this reason, a number of more robust (non-parametric) alternatives have been suggested. An example of these is given by Cliff (1993). There are also effect size

measures for multivariate outcomes. A detailed explanation can be found in Olejnik and Algina (2000). Finally, a method for calculating effect sizes within multilevel models has been proposed by Tymms et al. (1997). Good summaries of many of the different kinds of effect size measures that can be used and the relationships among them can be found in Snyder and Lawson (1993), Rosenthal (1994) and Kirk (1996).

Finally, a common effect size measure widely used in medicine is the 'odds ratio'. This is appropriate where an outcome is dichotomous: success or failure, a patient survives or does not. Explanations of the odds ratio can be found in a number of medical statistics texts, including Altman (1991), and in Fleiss (1994).

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 (with confidence intervals) 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 (i.e. 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 outcomes, different operationalisations of the same outcome, different treatments, or levels of the same treatment, or measures derived from different populations.
- The word 'effect' conveys an implication of causality, and the expression 'effect size' should therefore not be used unless this implication is intended and can be justified.

---

[1] 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 I. 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,

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

References

ALTMAN, D.G. (1991) *Practical Statistics for Medical Research*. London: Chapman and Hall.

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.

BAUGH, F. (2002) 'Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably'. *Educational and Psychological Measurement*, 62, 2, 254-263.

CLIFF, N. (1993) 'Dominance Statistics – ordinal analyses to answer ordinal questions' *Psychological Bulletin*, 114, 3. 494-509.

COHEN, J. (1969) *Statistical Power Analysis for the Behavioral Sciences*. NY: Academic Press.

COHEN, J. (1994) 'The Earth is Round (p<.05)'. *American Psychologist*, 49, 997-1003.

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.

DOWSON V. (2000) "Time of day effects in school-children's immediate and delayed recall of meaningful material". *TERSE Report* http://www.cem.dur.ac.uk/ebeuk/research/terse/library.htm

FINN, J.D. AND ACHILLES, C.M. (1990) 'Answers and questions about class size: A statewide experiment.' *American Educational Research Journal*, 27, 557-577.

FITZ-GIBBON C.T. (1984) 'Meta-analysis: an explication'. *British Educational Research Journal*, 10, 2, 135-144.

FITZ-GIBBON C.T. (2002) 'A Typology of Indicators for an Evaluation-Feedback Approach' in A.J.Visscher and R. Coe (Eds.) *School Improvement Through Performance Feedback*. Lisse: Swets and Zeitlinger.

FLEISS, J.L. (1994) 'Measures of Effect Size for Categorical Data' in H. Cooper and L.V. Hedges (Eds.), *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

FLETCHER-FLINN, C.M. AND GRAVATT, B. (1995) 'The efficacy of Computer Assisted Instruction (CAI): a meta-analysis.' *Journal of Educational Computing Research*, 12(3), 219-242.

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.

HARLOW, L.L., MULAIK, S.S. AND STEIGER, J.H. (Eds) (1997) *What if there were no significance tests?* Mahwah NJ: Erlbaum.

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.

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

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.

KESELMAN, H.J., HUBERTY, C.J., LIX, L.M., OLEJNIK, S. CRIBBIE, R.A., DONAHUE, B., KOWALCHUK, R.K., LOWMAN, L.L., PETOSKEY, M.D., KESELMAN, J.C. AND LEVIN, J.R. (1998) 'Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses'. *Review of Educational Research*, 68, 3, 350-386.

KIRK, R.E. (1996) 'Practical Significance: A concept whose time has come'. *Educational and Psychological Measurement*, 56, 5, 746-759.

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.

LEPPER, M.R., HENDERLONG, J., AND GINGRAS, I. (1999) 'Understanding the effects of extrinsic rewards on intrinsic motivation - Uses and abuses of meta-analysis: Comment on Deci, Koestner, and Ryan'. *Psychological Bulletin*, 125, 6, 669-676.

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.

LIPSEY, M.W. AND WILSON, D.B. (1993) 'The Efficacy of Psychological, Educational, and Behavioral Treatment: Confirmation from meta-analysis.' *American Psychologist*, 48, 12, 1181-1209.

MCGRAW, K.O. (1991) 'Problems with the BESD: a comment on Rosenthal's "How Are We Doing in Soft Psychology'. *American Psychologist*, 46, 1084-6.

MCGRAW, K.O. AND WONG, S.P. (1992) 'A Common Language Effect Size Statistic'. *Psychological Bulletin*, 111, 361-365.

MOSTELLER, F., LIGHT, R.J. AND SACHS, J.A. (1996) 'Sustained inquiry in education: lessons from skill grouping and class size.' *Harvard Educational Review*, 66, 797-842.

OAKES, M. (1986) *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. New York: Wiley.

OLEJNIK, S. AND ALGINA, J. (2000) 'Measures of Effect Size for Comparative Studies: Applications, Interpretations and Limitations.' *Contemporary Educational Psychology*, 25, 241-286.

ROSENTHAL, R. (1994) 'Parametric Measures of Effect Size' in H. Cooper and L.V. Hedges (Eds.), *The Handbook of Research Synthesis*. New York: Russell Sage Foundation.

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.

SNYDER, P. AND LAWSON, S. (1993) 'Evaluating Results Using Corrected and Uncorrected Effect Size Estimates'. *Journal of Experimental Education*, 61, 4, 334-349.

STRAHAN, R.F. (1991) 'Remarks on the Binomial Effect Size Display'. *American Psychologist*, 46, 1083-4.

THOMPSON, B. (1999) 'Common methodology mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap.' Invited address presented at the annual meeting of the American Educational Research Association, Montreal. [Accessed from <http://acs.tamu.edu/~bbt6147/aeraad99.htm>, January 2000]

TYMMS, P., MERRELL, C. AND HENDERSON, B. (1997) 'The First Year as School: A Quantitative Investigation of the Attainment and Progress of Pupils'. *Educational Research and Evaluation*, 3, 2, 101-118.

VINCENT, D. AND CRUMPLER, M. (1997) *British Spelling Test Series Manual 3X/Y*. Windsor: NFER-Nelson.

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

WILCOX, R.R. (1998) 'How many discoveries have been lost by ignoring modern statistical methods?'. *American Psychologist*, 53, 3, 300-314.

WILKINSON, L. AND TASK FORCE ON STATISTICAL INFERENCE, APA BOARD OF SCIENTIFIC AFFAIRS (1999) 'Statistical Methods in Psychology Journals: Guidelines and Explanations'. *American Psychologist*, 54, 8, 594-604.