What Effect Size Is and Why It Is Important

Robert Coe
School of Education, University of Durham

https://www.leeds.ac.uk/educol/documents/00002182.htm

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

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.

'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.

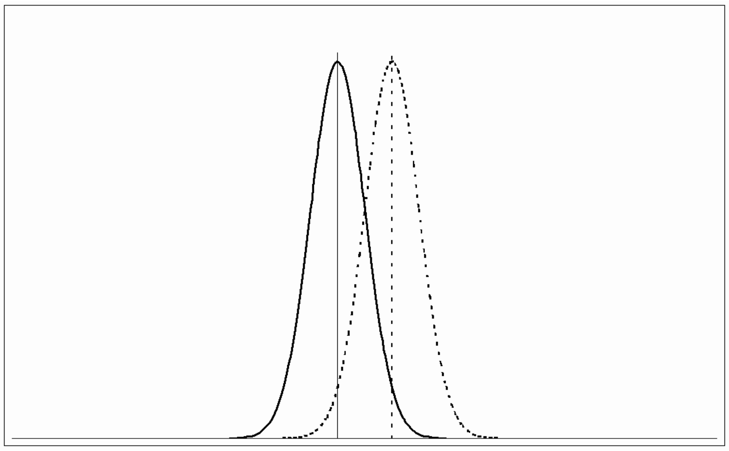

Figure 1a

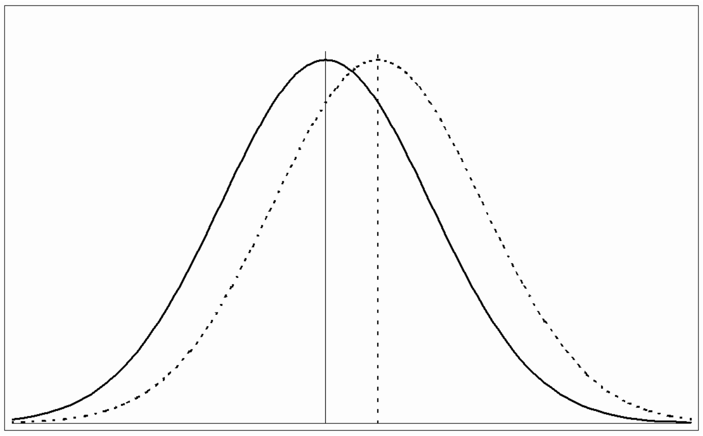

Figure 1b

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:

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

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.