But what does it *mean*?

The use of effect sizes in educational research

Edited by Ian Schagen and Karen Elliot

nfer

INSTITUTE OF
EDUCATION
UNIVERSITY OF LONDON

5 Effect sizes in multilevel models

Peter Tymms

5.1 Introduction

Several different approaches allow quantitative researchers to report the size of the effect being studied. When using multilevel models it has become common to discuss 'the proportion of variance accounted for' as well as 'the intra-class correlation'. These two measures combined with a direct interpretation of the coefficients can provide a clear picture. But there has been a growing interest in the use of effect sizes as used in experimental designs as a measure of the size of an effect and this paper explores their possible use within multilevel modelling.

By way of illustration the discussion is restricted to multilevel models found in educational research in which pupils are nested within schools. The model therefore has two levels. Before any explanatory variables are added the equations representing the null model are:

$$\text{At the pupil level:} \quad y_{ij} = \beta_{0j} + e_i \quad (1)$$

$$\text{At the school level:} \quad \beta_{0j} = \beta_0 + u_j \quad (2)$$

These may be combined to give a single equation:

$$y_{ij} = \beta_0 + u_j + e_i \quad (3)$$

Where:

y_{ij} is the outcome measure for pupil i in school j

β_{0j} is a constant which varies across schools

e_i is the error on the pupil measures

u_j is the error on the school measures

σ_e^2 is the variance at the pupil level

σ_u^2 is the variance at the school level

Effect sizes have been defined in relation to interventions in which there is a control and an experimental group. Glass *et al.* (1981) defined effect sizes as the difference between the mean scores for the experimental and control groups expressed in Standard Deviation (SD) units. The SD was taken to be that of the control group. More recently Hedges and Olkin (1985, p. 78) have argued that the pooled SD should be used rather than the SD of one particular group and that is now the more commonly accepted definition, which will be used in this paper, although it should be noted that Glass and Hopkins (1996, p. 290) still prefer the earlier version. The Hedges and Olkin version will be used in this paper and the formula is:

$$\Delta = \frac{\overline{X}_{Exp} - \overline{X}_{Cont}}{SD_{pooled}}$$

(4)

In other words the effect size is the difference between the means for the experimental and control groups expressed as a fraction of the pooled standard deviation.

This definition will be used to explore effect sizes in multilevel models under three headings. The first will look at dichotomous variables, the second at continuous variables, and the third at units that are conceived of as being measured on a continuous scale (random effects).

5.2 Where the variable is dichotomous

Suppose that some schools employed a psychologist and some did not. This may be represented by a dummy variable in the multilevel model and a coefficient associated with the variable is generated. Ideally the study would be an experimental one in which psychologists have been randomly assigned to schools, but it may also be that the controls are statistical. Ignoring any control variables for a moment the equation becomes:

$$y_{ij} = \beta_0 + \beta_1 + u_j + e_i$$

(5)

Where β_1 is the dummy variable representing the presence of a psychologist.

Now the calculation of the effect size is simply the difference in the means for the schools with and without psychologists (β_1) divided by the pooled standard deviation (the square root of the within group variance). This is simply σ_e; the standard deviation at the pupil level and the equation for the effect size is:

$$\Delta = \frac{\beta_1}{\sigma_e} \tag{6}$$

This formula and others in this section were first published in Tymms *et al.* (1997).

An example comes from the ESRC funded investigation (Tymms and Merrell, 2003) in which booklets were randomly assigned to schools. The booklets were designed to help teachers work with children who were inattentive, impulsive and hyperactive. The results of one very simple model of the data are given below:

Table 5.1 Outcome measure: attitude to reading (mean=-0.045 SD=0.88)

	Coefficients
Fixed	
Cons	–0.062 (0.013)
Dummy to indicate booklet	0.038 (0.020)
Random	
Pupil	0.739 (0.008)
School	0.029 (0.003)

The coefficient associated with the random assignment of the booklet was not statistically significant at the 5% level but it is still important to estimate the effect size since the coefficient is the best available evidence for the impact of the booklet. This is a quite different position from the stance which says that there was not effect, i.e. that the proper position is to stick to the null hypothesis, and this stance has been cogently argued for on numerous occasions (see for example Cahan, 2000).

The effect size from the model is $0.038/\sqrt{0.739} = 0.044$

The error on the effect size must be calculated by combining the errors from both the coefficient and the SD. If it is necessary to combine the errors then the general formula may be applied:

If the error in X is errX and X=A/B or A*B then:

$$\frac{errX}{X} = \sqrt{\left(\frac{errA}{A}\right)^2 + \left(\frac{errB}{B}\right)^2} \tag{7}$$

In this case the error on the coefficient is proportionally very much greater than the error on the SD (53% of 1%), which can therefore be ignored.

So the error can be set at 53%.

The effect size was 0.044 +/– 0.023

As noted above it has been assumed that the design was equivalent to an experimental design with no controls. Where multilevel models employ additional controls the pooled standard deviation of pupil scores σ_e drops. The question then arises as to whether the standard deviation before or after controls should be used in the calculation of the effect size. This depends on how one conceives of the experimental parallel. Let us suppose that the outcome measure was an attainment measures and the major control was prior achievement from a few years earlier. This will have resulted in a large drop in the pupil level variance of about a half and the SD therefore falls by about 70 per cent. If the effect size is now calculated using the reduced pupil level SD then this is parallel to an experimental design in which pupils of similar prior scores were selected to be part of the design and half were randomly assigning the treatment.

This is a perfectly proper experiment to do, but of course the standard deviation of the group will be somewhat less than if one had worked with the full range. So although it might seem unfair to use the final standard deviation (after controls), as long as one defines what one is doing then the standard deviation from the final model is appropriate.

The data on attitude to reading and the random assignment of booklet provides an example. When a control for the children's starting points was added the model became as shown in Table 5.2 below:

Table 5.2   Model with inclusion of control

	Coefficients
Fixed	
Cons	−0.075 (0.014)
Dummy to indicate booklet	0.053 (0.020)
Baseline	0.091 (0.008)
Random	
Pupil	0.734 (0.009)
School	0.029 (0.004)

Now the assignment of booklets is significant at the 5% level and the effect size is:

$$0.053/\sqrt{0.734} = 0.061$$

In this case the pupil level variance was hardly affected by the control variable but the coefficient associated with the dummy variable did change.

As an aside it is worth noting that the above discussion, concerning which SD should be used when calculating effect sizes, raises an issue for those engaged in meta-analyses since protocols in the standard procedures do not involve any coding of the primary investigations relating to the degree to which interventions were restricted to sub-samples of the population.

5.2.1 When the dummy variable is not a school effect

Variables often appear in multilevel models simply as controls. That is to say, they are there to improve the model or because there is an inherent interest in them and not because they measure school differences *per se*.

For example, a treatment might have been randomly assigned within schools but not across schools and there is an interest in the size of the effect, but it comes from a different perspective than that described above. Of course, it might be that the impact of the within school experiments varied across schools. The section below on units that form a continuum covers calculations of such effect sizes.

If the variable has been randomly assigned within schools then the SD used for the calculation of the effect size should not be σ_e but rather the pooled SD of the experimental and control groups. Such information does not appear in a basic multi-level model but can be obtained by fitting separate level 1 variances for the two groups. More details can be found in Rasbash *et al.* (1989, p. 18).

But although it is proper to run such models and to carry out the calculation to produce an unbiased estimate of the effect size if the effect size is small the result will be almost the identical to that produced using σ_e. The question is: how small is small? The chart below helps to quantify the answer. It shows the results of a simulation using 10,000 cases and it suggests that if effect sizes were estimated to be 0.4 or lower then no advantage is to be had in calculating effect sizes by more complex analyses than using the formula β / σ_e. However, if it was greater than 0.4 then the effect size will be underestimated by an educationally important amount. An effect size of 1 will appear to be a little more than 10% lower than the true value.

Figure 5.1 Effect sizes calculated using σ_e and the pooled SD

ES using gross SD (incorrect)

ES using pooled SD (correct)
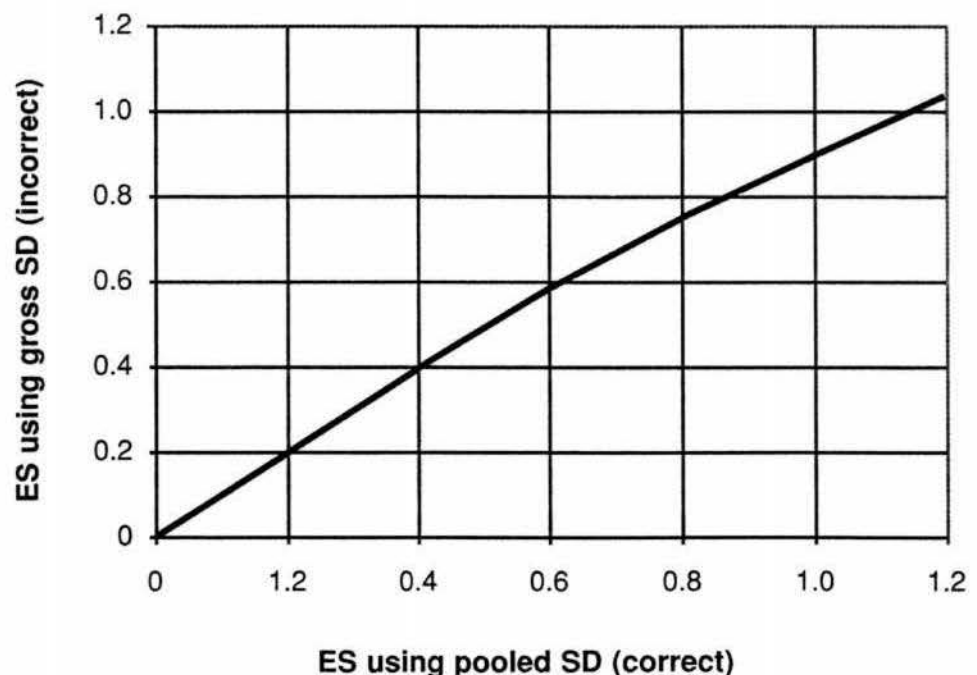
5.3 Where the measure is continuous

It may be that a measure thought to impact on schools forms a continuous variable and this may have been randomly assigned to schools. Varying amounts of inspection time, for example, may have been allocated to schools. When a continuous variable is employed the parallel from Glass *et al.* (1981) is a correlation and they suggest:

$$\Delta = 2z r_{xy}(1 - r_{xy}^2)^{-\frac{1}{2}} \tag{8}$$

where:

r is the correlation between variables x and y

z is the 'unit normal deviate at the pth percentile'

Extracting an effect size from a continuous variable involves considering it as though it were a dichotomous variable and deciding where to slice the continuous variable. If this is chosen as one SD above and below the mean then this simplifies according to Fitz-Gibbon and Morris (1987) to:

$$ES = \frac{2r}{\sqrt{(1 - r^2)}} \tag{9}$$

This is equivalent to the difference between the residuals of the standardised criterion corresponding to predictor scores one SD above and one SD below the mean expressed as a fraction of the SD of the residuals. This equation can be 'seen' in Figure 5.2, which shows the scatterplot of two normally distributed variables each with a mean of 0 and a SD of 1. The slope of the line is equal to the correlation coefficient (r). Vertical lines have been drawn from the mean on the x-axis and from point one SD above and below the mean. Horizontal lines are then drawn from the points where these lines meet the regression line to the y axis and the effect size is the distance between the points marked r and –r divided by the SD of the residuals from the regression:

Figure 5.2 Graphical representation of the effect size using a continuous variable
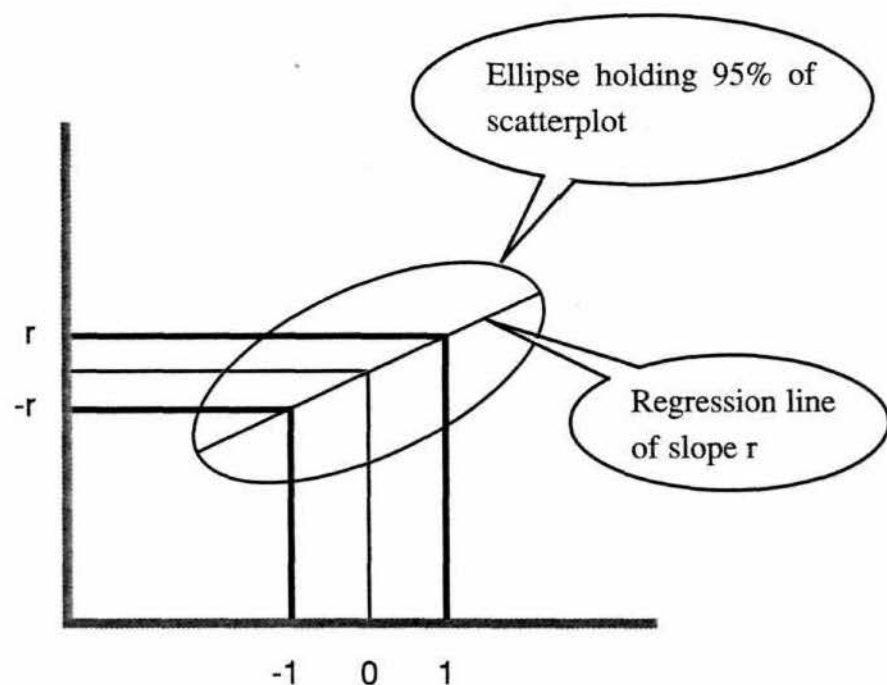
In a simple multilevel model in which the continuous predictor and outcome variables have been normalised (mean = 0; SD = 1) the coefficient is equivalent to r and the standard deviation of the pupil level scores, after controls, is σ_e. The formula for effect size becomes:

$$\Delta = \frac{2\beta_1}{\sigma_e} \tag{10}$$

A slightly more complex formula is required if the predictor and criterion are not z scores. Consider Figure 5.2. The slope of the line is now β_1 and the positions of the vertical lines correspond to one $SD_{predictor}$ to the right and left of the mean. Hence the distance between what was r and −r becomes $2\,\beta_1 * SD_{predictor}$. The formula is:

$$\Delta = \frac{2\beta_1 * SD_{predictor}}{\sigma_e} \tag{11}$$

Using the multilevel model in the last box the effect size for the main control (baseline) can be calculated given the SD of the baseline measure which is 1.

The effect size is:

$2*0.091* 1 / \sqrt{0.734}$

or 0.21

As in the last section the same discussion relating to the presence of control variables in the model and the impact that that has on the value of σ_e applies.

5.4 There are units (schools) that form a continuum

In this case a similar approach can be used and now the distance between one standard deviation above to one standard deviation below is twice the standard deviation at the school level and the formula is straightforward:

$$\Delta = \frac{2\sigma_u}{\sigma_e} \qquad (12)$$

Again no account is taken of explanatory variables and the same argument applies as appeared earlier.

Using the last multilevel model, the effect size for the school effect can readily be calculated. It is:

$2*\sqrt{0.029} / \sqrt{0.734}$

or 0.40

This is a measure of the importance of the school in children's attitudes to reading.

5.5 Relationship of effect size to r² and to the intra-class correlation

A general measure of the magnitude of a regression coefficient is the proportion of variance 'explained' by its inclusion in the equation. This is equal to the squared correlation coefficient. Hedges and Olkin (1985, p. 77) state that for equal sized experimental and control groups the link between the two measures (proportion of variance and effect size) is:

$$\rho^2 = \frac{\Delta^2}{\Delta^2 + 4} \tag{13}$$

where:

ρ is the correlation coefficient

Δ is the effect size

This equation can be rearranged to give the formula quoted from Fitz-Gibbon and Morris (1987) earlier and gives a clear link between the proportion of variance 'explained' and effect size. This is shown diagrammatically in Figure 5.3.

Figure 5.3 The link between the proportion of variance 'explained' and effect size
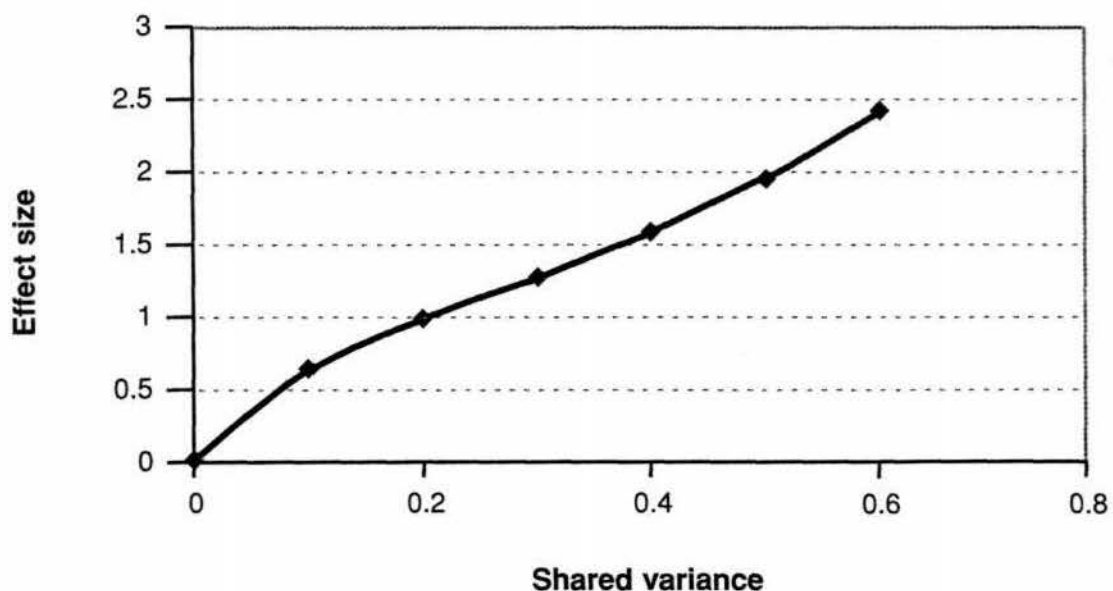
It is common practice to express the size of the school effect in terms of the proportion of variance associated with the school. This is the intra-
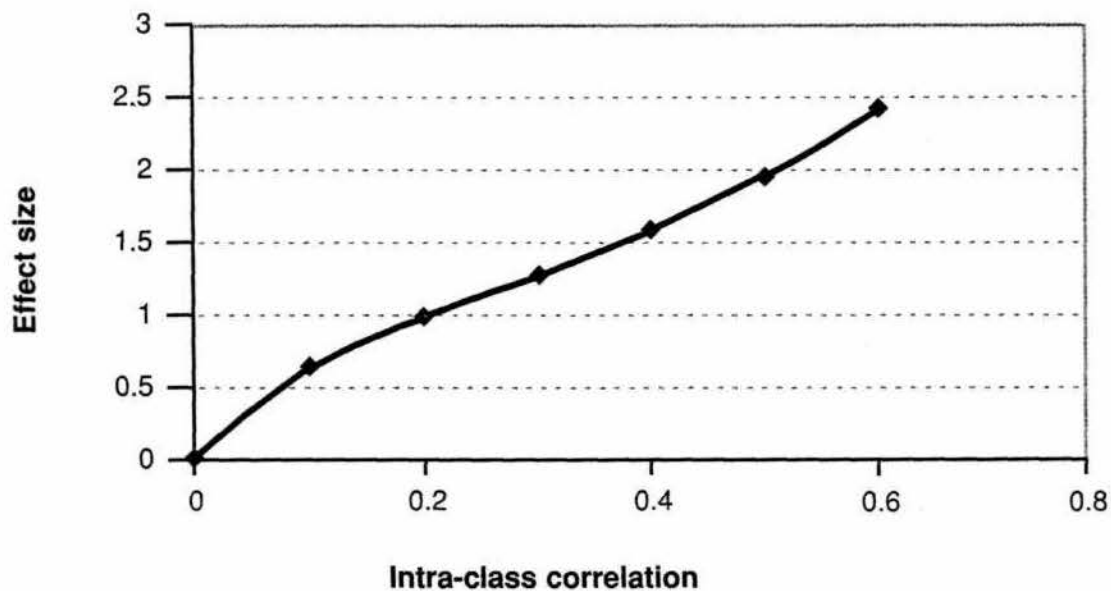
class correlation (ρ) and is given by:

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \tag{14}$$

When the earlier formula expressing the effect size in terms of and is combined with the above it gives:

$$\Delta = \sqrt{\frac{4\rho}{1-\rho}} \tag{15}$$

The relationship is shown in Figure 5.4.

Figure 5.4   Relationship between effect size and intra-class correlation

N.B. The similarity between Figures 5.3 and 5.4 arises because the rho in the intra-class correlation formula is the proportion of variance and this parallels r^2 in the earlier effect size formula.

5.6 Conclusion

This paper has set out a straightforward way of addressing the issue of effect sizes when using multilevel models to study schools. It has provided formulae that allow effect sizes to be calculated in standard

deviation units and has shown how these relate to the more commonly used measures of the sizes of effects in multilevel modelling, which are expressed in alternative forms. The effect sizes in multilevel models have been conceptualised in experimental terms so that there can be a clear understanding of what they mean.

The paper has not addressed issues associated with non-normal distributions, non-linear relationships nor has it dealt with anything other than very simple multilevel models.

References

CAHAN, S. (2000). 'Statistical significance is not a "Kosher Certificate" for observed effects: a critical analysis of the two-step approach to the evaluation of empirical results', *Educational Researcher*, **29**, 1, 31–4.

FITZ-GIBBON, C.T. and MORRIS, L.L. (1987). *How to Analyze Data*. Beverly Hills, CA: Sage.

GLASS, G.V. and HOPKINS, K.D. (1996). *Statistical Methods in Education and Psychology*. Third edn. Boston, MA: Allyn and Bacon.

GLASS, G.V., McGAW, B. and SMITH, M.L. (1981). *Meta-analysis in Social Research*. London: Sage.

HEDGES, L.V. and OLKIN, I. (1985). *Statistical Methods for Meta-analysis*. New York, NY: Academic Press.

RASBASH, J., PROSSER, R. and GOLDSTEIN, H. (1989). *ML2 Software for Two-Level Analysis*. London: University of London, Institute of Education.

TYMMS, P., MERRELL, C. and HENDERSON, B. (1997). 'The first year at school: a quantitative investigation of the attainment and progress of pupils', *Educational Research and Evaluation*, **3**, 2, 101–18.

TYMMS, P.B. and MERRELL, C. (2003). 'Screening and interventions for inattentive, hyperactive and impulsive children.' Paper presented at the Evidence Based Policy and Practice Conference, London, July.