On 'But what does it mean?': the use of effect sizes in educational Research
Editors: I. Schagen & K. Elliot. NFER, Slough 2004.

6 Some observations on the definition and estimation of effect sizes

Harvey Goldstein

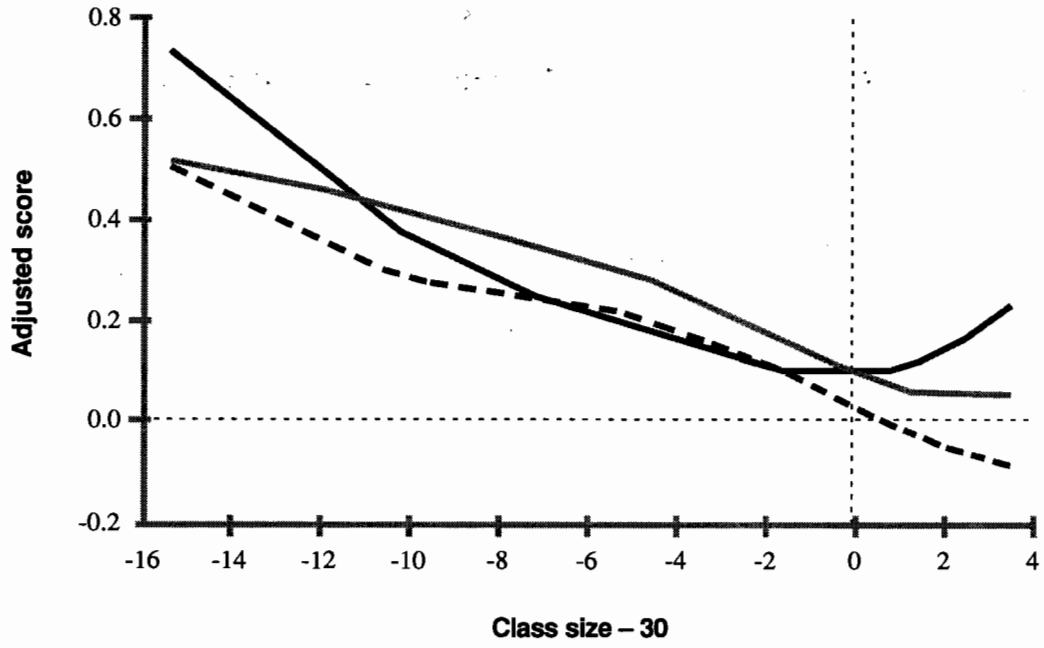
6.1 General considerations

Many valuable comments have been made by contributors and while some of the key issues have been aired, I would like to suggest that prior to considering effect sizes it is important to pay attention to the correct specification of the statistical model being used. Thus, for standard regression or multilevel models the assumption of Normality is typically made and much of the literature on effect sizes, especially that which concentrates on standardised effects, assumes Normality. A prior transformation to Normality, for example using Normal scores, may often be needed for both the response and predictor distributions. Likewise, the existence of complexity in the form of interactions, or random coefficients in a multilevel model, should be explored and where such complexities exist a graphical presentation of effects will usually be especially helpful.

The most common reason for wishing to use standardised effect sizes is to compare findings from different studies, as in meta analyses. Where comparisons are made between explanatory (predictor) variable coefficients in the same model, some care is needed since these explanatory variables and the coefficient estimates may be highly correlated. In any case it is good practice to estimate a confidence interval for the difference between two such standardised coefficients, or carry out a test of significance.

A particular important case is where the relationship between a response and a predictor variable is non-linear so that a simple effect size in the form of a standardised regression coefficient is unavailable. In a recent study of class size effects (Blatchford *et al.*, 2002) not only was the relationship between test score (adjusted for prior attainment) and class size non-linear, there were also interactions between this relationship and level of prior attainment. Figure 6.1 presents these relationships in a way that shows clearly what is occurring. It would be difficult to find a simple alternative method of presentation using effect size estimates.

Figure 6.1 Reception literacy by class size for three baseline groups



The response is a literacy test score taken at the end of reception year and adjusted for the prior baseline test score and other factors; the line with the steepest slope for class sizes below about 23 is that for the lowest achieving group at entry to reception class. The non-linearities are important since they illustrate the changing relationship for this group for class sizes over about 27. The model was fitted using cubic regression splines within a multilevel model and is an interesting example of where traditional methods of fitting linear relationships and quoting effect sizes based upon the resulting regression coefficients would have presented a distorted view of the underlying reality.

In the remainder of this contribution I will comment on the following specific issues. The first is the question of the appropriate units in which to present results and how to form a standardised coefficient. The second is how one might deal with binary (or ordered) predictor and response variables and finally I will make some comments on the use of utility or cost functions for comparing 'effects'.

6.2 Presentation and units of reporting

In a simple linear regression model one can form a standardised regression coefficient which will denote the estimated change in

standard deviation units of the response for a change in 1 standard deviation of the predictor. Whether or not one chooses the response distribution before or after fitting the predictor variable (i.e. based on the residual variance) will depend on purpose. For example, if the model is a multilevel one and includes school class as a random factor and the predictor of interest is measured at the class level, say class size, then the within class level 1 residual variance would seem to be the appropriate one to use, since this is more likely to be comparable across studies since these may have very different percentages of relative between-classroom variance. On the other hand, if the predictor of interest is measured at the individual level then the overall population standard deviation would seem to be more appropriate for purposes of reporting and comparing effects. In a randomised controlled trial where treatments are administered to individuals the use of the control group S.D. reflects this, since that is the naturally occurring S.D. in the population.

The ideal situation is where there is a 'natural' reporting unit. In education, with young children this might be years of progress associated with the response measure that is reporting an effect in terms of the average years of progress for a unit change in the standardised predictor. Blatchford *et al.*, (2002) use this, but remark that the conversion of score scales to years of progress requires data from longitudinal studies that are usually not available. The age standardisations typically supplied by test publishers are in fact a mixture of 'cross sectional' and 'longitudinal' adjustments that are not suitable (see Goldstein and Fogelman, 1974 for a further discussion). Another possibility is to choose a standard metric against which other effects will be calibrated. Thus, we might choose the girl–boy difference, suitably contextualised for age and response type, and present other effects as multiples of this.

6.3 Binary variables

The first case is where we have a binary response variable, say a pass/fail indicator, rather than a continuous score. A standard statistical procedure is to assume an underlying continuous distribution which has a threshold above which the indicator (say an exam pass) is triggered. A probit analysis can be carried out where the underlying continuous distribution is assumed to be a standard Normal one and this then allows direct calculation of a standardised regression coefficient. Where the response is ordered, for example a 5-point scale, then a similar procedure can be

implemented. For comparability purposes of reporting effect sizes and being able to compare with continuous response variable analyses, such analyses should be carried out in preference to the more common logit modelling – although the general statistical inferences concerning significance etc. will generally be little changed.

The second case is the one discussed by Schagen (in Chapter 3) where we have a binary predictor. In such cases, we need to distinguish between cases where it is reasonable to assume an underlying continuum such as, say, social status and where there is no such concept as in the case of gender or type of school. Where there is no reasonable assumption of an underlying continuum it just does not seem appropriate to attempt to define an effect size that is comparable to one defined for a continuous variable and I do not see that any amount of mathematical manipulation is appropriate in such cases. Where we can assume an underlying continuum then the following simple approach suggests itself.

Suppose the predictor is social class measured as manual/non-manual and we assume an underlying social status continuum. As a simple illustration, suppose that the proportion manual is 0.5 and suppose also that in a simple analysis, using a standardised (or Normalised) response, for the binary social class variable the social class difference is estimated to be 0.2 units – i.e. this is the coefficient of the dummy variable for social class. Using the probit idea described above we suppose that there is an underlying standard Normal distribution where the mean of zero in this case corresponds to the cut-off between manual and non-manual, since the proportion of manual is 0.5. If we assume that those with a manual social class are randomly sampled from the underlying distribution then their average value from this distribution is simply the average for the Normal distribution truncated above at zero, which is about –0.8. Likewise the non-manuals will have an average on the underlying distribution of about 0.8.

Thus, the difference on the underlying normal is 1.6 units, rather than the 1.0 units implied by using a standard dummy variable coding. Therefore, if we divide the estimate above of 0.2 by 1.6 to give 0.13 we have an estimate for the coefficient that we would have if we actually used a direct measure of the underlying social status having a standard Normal distribution; this will be the effect size. It is possible to extend this idea to ordered categories, but it does rest upon the assumption that, given the

category, e.g. manual, there is no association between the underlying continuous distribution values and any other predictor variables, and in general we might not expect this to be true.

A more sophisticated approach to this problem will take account of this possibility and Gibbs sampling (Albert and Chibb, 1993) can be used for the estimation. Research on this, with a view to incorporating it into MLwiN (see Browne, 2003) is currently being pursued.

6.4 Utilities and costs

Instead of attempting to provide single number summary comparisons for different variables that can be compared across studies, it might be better to give the user responsibility for deciding how to make such comparisons. Suppose we have two predictors, a measure of special educational need (yes/no) and gender. We can ask the user of our analysis to place relative costs on having a gender difference and having a difference between our special educational needs groups. Such costs might be thought of in terms of the social utility of eliminating such differences or perhaps the resource costs of doing so, or some combination. Suppose that the estimated difference between categories in our model is the same for both variables but the utility for special needs is thought to be twice that for gender. This would imply that eliminating the category of children with special needs will result in a greater (twice) social 'gain' than eliminating the gender difference, and this might then guide policy.

Of course, this is only a crude example and all kinds of objections can be raised, but allowing considerations of utility and cost to enter at the stage of presenting results, as a product of discussions with users, does seem to have something to recommend it and avoids at least some of the drawbacks associated with presenting users with single estimates of effect sizes.

References

ALBERT, J.H. and CHIB, S. (1993). 'Bayesian analysis of binary and polychotomous response data', *Journal of the American Statistical Association*, **88**, 669–79.

BLATCHFORD, P., GOLDSTEIN, H., MARTIN, C. and BROWNE, W. (2002). 'A study of class size effects in English school reception year classes', *British Educational Research Journal*, **28**, 2, 169–85.

BROWNE, W.J. (2003). *MCMC Estimation in MLwiN*. London: University of London, Institute of Education.

GOLDSTEIN, H. and FOGELMAN, K. (1974). 'Age standardisation and seasonal effects in mental testing', *British Journal of Mathematical and Statistical Psychology*, **44**, 109–15.