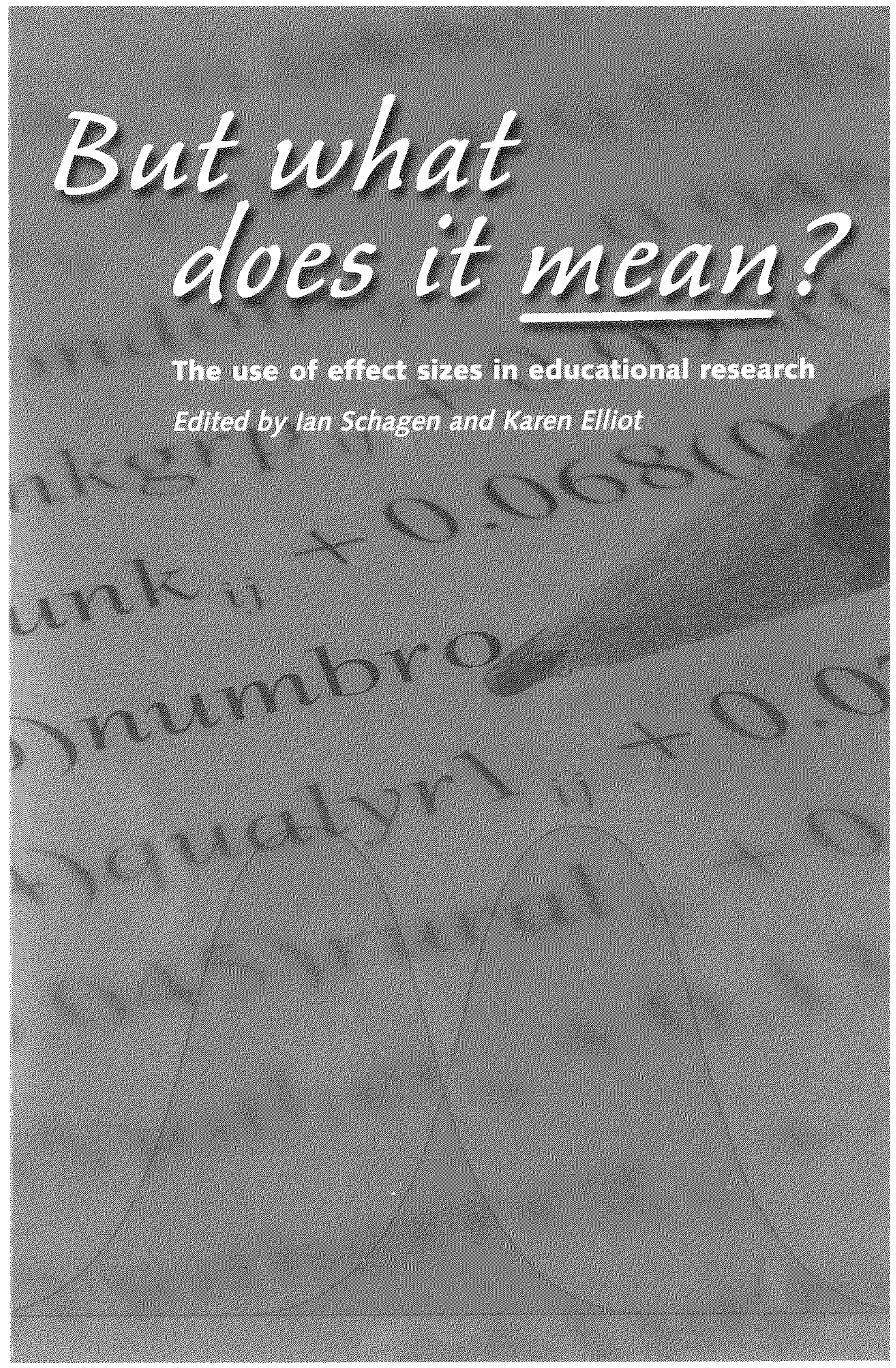


But what does it *mean*?

The use of effect sizes in educational research

Edited by Ian Schagen and Karen Elliot

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

First published in 2004 by the National Foundation for Educational Research,
The Mere, Upton Park, Slough, Berks SL1 2DQ
www.nfer.ac.uk

ISBN 1 903880 67 X

A catalogue record for this book is available from the British Library.

Typeset by Mary Hargreaves, National Foundation for Educational Research
Cover design by Stuart Gordon, National Foundation for Educational Research
Printed in Great Britain by B & B Press, Rotherham

But what does it *mean*?

The use of effect sizes in educational research

Contents

Contributors

Robert Coe
CEM Centre, University of Durham

Karen Elliot
Institute of Education University of London

Ray Godfrey
Canterbury Christ Church University College, Canterbury

Harvey Goldstein
Institute of Education University of London

Michela Gnaldi
Statistics Research and Analysis Group,
National Foundation for Educational Research, Slough

John Gray
Faculty of Education, University of Cambridge

Paula Hammond
Statistics Research and Analysis Group,
National Foundation for Educational Research, Slough

Trevor Knight
Analytical Services, Department for Education and Skills, London

Pam Sammons
Institute of Education University of London

Ian Schagen
Statistics Research and Analysis Group,
National Foundation for Educational Research, Slough

Caroline Sharp
Professional and Curriculum Studies,
National Foundation for Educational Research, Slough

Steve Strand
nferNelson, London

Peter Tymms
CEM Centre, University of Durham

Introduction

Karen Elliot and Ian Schagen

But what does it *mean*? This is a question that is increasingly being asked, in particular by policy makers, to researchers and academics when presenting research findings. An exploration of the use of effect sizes within educational research may provide a first step towards answering that question.

It was within this context that, in November 2003, an invitational seminar was jointly organised by the Institute of Education University of London and the National Foundation for Educational Research (NFER). The aim of the seminar was to provide a forum for researchers, academics and policy makers to debate the issues surrounding the calculation and interpretation of effect sizes in educational research and, more specifically, school effectiveness research (SER) employing, for example, multilevel modelling.

The emphasis in the morning seminar session was on the use of effect sizes within complex statistical methodologies with four papers presented on this theme and two discussants providing responses. The aim of the shorter afternoon session (with two papers and one discussant) was to place the issue of effect sizes into a wider context within the educational landscape. Throughout the day, the Chair highlighted the important connections between these dual themes, providing interesting anecdotes from his own prolific research experience and expertly steering the informed discussion between presenters and invited delegates.

It is hoped that the seminar and this publication of the day's proceedings will encourage researchers and academics to use tools such as effect sizes to facilitate the dissemination of research findings to stakeholders in the data, such as policy makers and school and LEA staff. A range of issues relating to this keenly debated theme have been raised, such as formulae to be used in the effect size calculation, the use of confidence intervals, model specification, presentation and interpretation of effect sizes. As a result, we believe that the debate on the use of effect sizes in educational research in the UK context has advanced and awareness has been increased, in particular with respect to effect sizes calculated from multilevel models. These advances, both theoretical and practical, can only further improve understanding for all involved and provide a significant step towards addressing the initial question of 'But what does it *mean*?'.

The first chapter is written by the Chair, John Gray, University of Cambridge, whose introductory comments and valuable insights set the papers and resulting discussion in the context of educational research, and in particular SER, over the past decades.

Chapter 2 written by Karen Elliot and Pam Sammons from the Institute of Education University of London, provides a background to the use of effect sizes. The notion of effect sizes is not a new concept, as a number of different types of formulae for the calculation of effect sizes have existed for many years. It is noted that, although a comprehensive number of appropriate statistics are detailed in SER, effect sizes are generally not routinely reported. Formulae are presented for calculating effect sizes in studies that investigate, using multilevel modelling, naturally occurring variation in child outcomes, with the possibilities and limitations of such a statistic also explored.

Ian Schagen of the NFER has contributed Chapter 3. He argues that the 'But what does it *mean*?' problem has partly arisen due to the ability to present and interpret research findings not keeping pace with advances in the capacity and power of statistical analyses. He discusses, amongst other ideas, the concept of dimensionless normalised coefficients and star wars plots (essentially a graphical method for presenting these measures and their confidence intervals).

Chapter 4 is contributed by Steve Strand from nferNelson who argues the usefulness of effect size measures as a supplement to statistical significance testing. Indeed, in certain contexts outlined by Strand, he considers that traditional significance testing is limited or insufficient. He also broadens the discussion of effect size beyond measures of central tendency by introducing the concept of the variance ratio to assess the magnitude of group differences in score variance.

Peter Tymms, from the University of Durham, furthers in Chapter 5 the exploration of effect sizes in multilevel models. A theoretical perspective is taken in which approaches and formulae for effect size calculation, which are intended to be of practical use, are presented for three situations: dichotomous variables, continuous variables and units that are conceived of as being measured on a continuous scale (random effects). He discusses the issue of which standard deviation to employ in the calculation of an effect size and explores the relationship between effect sizes and other statistics commonly reported in SER studies.

Chapters 6 and 7 consist of the comments from the two discussants from the morning session of the seminar linked to the use of effect sizes within complex methodologies, relating to Chapters 2–5 of this volume. In Chapter 6, Harvey Goldstein of the Institute of Education University of London, provides some observations on the definition and estimation of effect sizes. He offers some precautionary advice to consider prior to calculating effect sizes and also warns that, in certain cases, quoting effect sizes based upon regression coefficients presents a distorted view of the underlying reality. The presentation and units of reporting are discussed as are binary predictor and response variables. Goldstein concludes with his thoughts on the use of utility and cost functions for comparing effects.

In Chapter 7 Trevor Knight from the DfES, in his role of discussant, gives a comprehensive summary of the expansion of educational data in the UK context and highlights the necessity to present results describing complex relationships in ways that are comprehensible to the intended audience.

Robert Coe from the University of Durham has written Chapter 8. The case for using effect size measures is presented with Coe arguing that effect size enables uncalibrated measures to be interpreted, emphasises amounts (not just statistical significance), draws attention to the margin of error, may help to reduce reporting bias and allows the accumulation of knowledge within a 'meta-analysis'. However, there are also a number of problems and complexities associated with the calculation and interpretation of effect sizes. Therefore, the author provides a list of recommendations to encourage good practice in the use and reporting of effect sizes within educational research.

A philosophical approach to the notion of effect size is presented in Chapter 9 by Ray Godfrey, Canterbury Christ Church University College. The meaning of a pseudo-concept is explored first with examples from the grammar of a language, then mathematics and statistics. Godfrey argues that effect sizes, rather than providing an answer to the question 'But what does it *mean*?' which has both sense and reference to the real world, offer a statistician's pseudo-concept.

In Chapter 10, Caroline Sharp from the NFER contributes her comments as discussant to the afternoon session (referring to Chapters 8 and 9 in this volume). Her stance is that of a non-statistician who, due to a close working relationship with statisticians, is exposed to such concepts as effect size. She has expertly synthesised the two papers, highlighting the commonalities between the two authors' perspectives as well as the different positions adopted.

Paula Hammond and Michela Gnaldi, also from the NFER, have written Chapter 11 as a synopsis of comments received from the floor at the seminar and subsequent remarks posted on the web discussion forum. They discuss both the technical and non-technical issues raised, clearly reporting the various points of view.

Finally, in Chapter 12 the editors provide a short summary about the use of effect sizes in educational research with particular reference to complex methodologies with some concluding comments to the 'But what does it *mean*?' debate.

Acknowledgements

The editors were delighted with the response to the seminar on effect sizes and genuinely surprised at the number of people within educational research who felt this was a sufficiently important topic to participate.

We would like to thank all those who contributed towards ensuring the success of the seminar and the book, including the Chair, presenters, discussants and those who contributed from the floor. We are extremely grateful to our two institutions (NFER and the Institute of Education University of London) for supporting and resourcing the event. Particular thanks to Sally Green, Paula Hammond, Michela Gnaldi, Kim Halliday and Wesley Welcomme for their hard work behind the scenes before and during the seminar. Paula and Michela have also been involved in the publication of the book and we appreciate all the efforts of NFER's communications and marketing staff in putting the publication together, especially Alison Lawson and Mary Hargreaves, and Stuart Gordon for designing the cover.

Both the seminar and this publication have been an exciting partnership between educational institutions and represent an outstanding team effort with which we are proud to have been associated.

The views expressed in the book are those of the contributing authors and are not necessarily the views of the NFER or the Institute of Education University of London.

1 Getting to grips with 'effect sizes': some reflections on a personal journey

John Gray

Looking back over three decades of experience as an educational researcher I can see that I have been grappling with the problem of 'effect sizes' for most of my career. For much of this time, however, the reporting of research findings has been fairly undisciplined.

My first encounter with something comparable to the calculation of 'effect sizes' came when I worked as a research assistant on Jencks' study of *Inequality* (Jencks *et al.* 1972). In reporting research conclusions all findings were converted into a series of top fifth/bottom fifth comparisons. Commenting on the influence of high schools on pupils' development, for example, Jencks concluded:

> *Overall, the evidence shows that differences between high schools contribute almost nothing to the overall levels of cognitive inequality. Differences between elementary schools may be somewhat more important but evidence for this is still somewhat inconclusive.* The average effect of attending the best rather than the worst fifth of all elementary schools is almost certainly no more than ten (standardised) points and probably no more than five. *The difference between, say, the top and bottom halves is even less* [my emphasis].
>
> (Jencks *et al.*, 1972, p. 93)

Seven years later, Rutter's *Fifteen Thousand Hours* was hailed as overturning the 'pessimism' of Jencks' analysis. Rutter wrote:

> *The school results range from 71.2% better than expectation to 55.4% below expectation – large differences indeed. The exam score, after adjusting for VR, of the most successful school (2.38) was nearly four times as high as that of the least successful (0.62).*
>
> (Rutter *et al.*, 1979, p. 86)

A decade later Smith and Tomlinson (1989) were reporting in similar vein:

> *The results of the present study show that there are very important differences between urban comprehensive schools. The level of*

> *achievement is radically higher in some schools than in others. The findings show that the same child would get a CSE Grade 3 in English at one school but an O level grade B in English at another. There are equally large differences in maths and in exam results in total across all subjects.*
>
> (Smith and Tomlinson, 1989, p. 301)

On first reading, then, the British research on school effectiveness seemed to contradict its American counterparts. Imagine my surprise, therefore, when I applied the Jencksian metrics to the two British analyses. The top fifth/bottom fifth estimates were virtually identical. The British researchers had simply emphasised the full range between the most and least effective schools in their (relatively small) samples without reference to their locations in the underlying distributions. The three studies were, I concluded, producing very similar estimates. The differences lay in their interpretation.

Of course, charges of over-claiming are not limited to the field of school effectiveness. Examples of 'over-interpreted' claims abound in educational research. Take the case of the claims for pre-schooling. American researchers have claimed that, in certain circumstances, for every dollar spent on pre-schooling society will ultimately save seven dollars on the costs of unemployment benefits, reduced crime, improved health and so on. It's a powerful claim but, unfortunately, not one that stands up to translation across time and context.

The pre-schooling programmes that have been implemented in the United Kingdom have mostly been very different indeed to the American pioneers. Furthermore, the Head Start Planned Variation Study, which formed the starting point for some of the outstandingly successful programmes, showed that what it termed the 'English infant school' model was mostly notable for having few effects either way on pupils' progress – positive or negative. A more realistic assessment of what pre-schooling in this country can deliver, rendered in terms of effect sizes, would have been helpful.

Within the last two years, in the debate about top-up fees, seemingly over-inflated claims have also been made. A university graduate apparently earns up to £400,000 more than a non-graduate over the course of their lifetime. This conclusion may be true for some graduates entering relatively lucrative professions, but important differences between academic subjects and the very restricted size of the graduate populations, on which the estimates have been based, have both somehow been forgotten. Again, the discipline of careful and measured comparison seems to have been largely ignored.

A determination to undertake cross-study comparisons in terms of a common framework does not, unfortunately, provide sufficient basis for proceeding as I discovered to my cost when I attempted to undertake the first British meta-analysis of the effects of class size. Stimulated by Glass's meta-analysis (Glass and Smith, 1979) but, at the same time, irritated that it ignored nearly all the British evidence, I set about identifying eleven relevant studies. With a little informed guesswork where certain details were not provided, I thought I was in a position to provide a reasonable estimate. Following recommended procedures for meta-analyses, however, I first asked a series of related questions, one of which was about the 'quality' of the research design on which the empirical estimates were based. To my surprise I noticed that in the process of undertaking the review I had coded all the studies available at that time as having 'low quality' research designs. End of meta-analysis!

The experience reinforced my conviction that the findings of any piece of research are only as good as the strategies that generated them, regardless of the frequency with which they have apparently been replicated. In a classic article some three decades ago about the problems of 'accumulating evidence' across sometimes rather disparate research studies, Light and Smith (1971) argued that it is improbable that one will actually find one single really well-designed study. What one needs to do is form a clear view of what a 'defensible' study would look like in the field under investigation before getting down to the business of comparing statistics across studies. In many cases this process of screening studies against set criteria will dramatically reduce the number of worthwhile contributions. But rather statistical estimates based on firm ground, Light and Smith would argue, that any number based on potentially shifting sands. 'The lesson we believe flows from these examples', they wrote, 'is that little headway can be made by pooling the *words in the conclusions* of a set of studies' (p.43). There is really no substitute for well-designed, implemented and analysed studies from which the statistical evidence is pooled. We ignore such advice at our peril.

Studies chosen for further analysis should, Light and Smith maintained, meet at least three standards. First, all subjects in the study should have been selected 'from a known and precisely definable population'. Second, a study's dependent variables and those independent variables that are measured 'must be measured in the same way as... those employed in the rest of the studies'. And third, 'the instrumentation and quality of the experimental work in a study must be generally comparable to that in all the rest of the studies' (pp.448–9). These strictures remain, in my view, as apposite today as they were then, even though our capacity to undertake sophisticated statistical modelling has dramatically increased since the time when they were writing.

So what is the way forward? That colleagues are ready to produce, research and debate the conceptual and technical issues surrounding the use and computation of 'effect sizes' represents an important step forward. Effect sizes, appropriately calculated, can help to impose order where previously indiscipline has reigned. However, as the contributions to this symposium underline, this is not just or simply a dry 'technical' matter. Different researchers favour different procedures. There is, as yet, no standard default model to which one can turn. Crucial choices need to be made.

Over the next few years the development of statistical understanding needs to go hand in hand with wider debates about how to improve the general qualities of educational research – in terms of study design and implementation, statistical analysis and, crucially, replication. We should not under-estimate the implications of attempting to produce more valid and reliable estimates of effect sizes. But we need to be cautious about leaping to the conclusion that larger is necessarily better. Sizeable effect sizes we can do little or nothing to modify are unlikely to contribute much to the cause of educational improvement. Indeed, such claims need, I would argue on the basis of experience, to be treated with a degree of scepticism.

As educationists become more discerning in their understanding of the routes to further improvement, the prominence given to comparatively modest but well-founded 'effect sizes' should increase. It is a testimony to the increasing maturity of educational research in this country that we have reached the point where informed discussion of the possibilities can begin.

References

GLASS, G. and SMITH, M. (1979). 'Meta-analysis of research on class size and achievement', *Educational Evaluation and Policy Analysis*, 2–16.

JENCKS, C., SMITH, M., ACLAND, H., BANE, M., COHEN, D., GINTIS, H., HEYNS, B. and MICHELSON, S. (1972). *Inequality: a Reassessment of the Effects of Family and Schooling in America.* London: Penguin Books.

LIGHT, R.J. and SMITH, P.V. (1971). 'Accumulating evidence: procedures for resolving contradictions among different research studies', *Harvard Educational Review*, **41**, 4, 429–71.

RUTTER, M., MAUGHAN, B., MORTIMORE, P. and OUSTON, J. (1979). *Fifteen Thousand Hours: Secondary Schools and their Effects on Children.* London: Open Books.

SMITH, D. and TOMLINSON, S. (1989). *The School Effect: a Study of Multi-Racial Comprehensives.* London: Policy Studies Institute.

2 Exploring the use of effect sizes to evaluate the impact of different influences on child outcomes: possibilities and limitations

Karen Elliot and Pam Sammons

2.1 Introduction

In today's data driven society, it is increasingly important to consider when presenting research findings how the various users and stakeholders employ the data. In the context of pupil performance, data relating to children's educational outcomes, a literature base is emerging linked to how schools (and also teachers, governors, parents, etc) use such feedback (Kluger and DeNisi, 1996; Dudley, 1999; Yang *et al.*, 1999; Saunders, 2000; Demie, 2003; Elliot and Sammons, 2001, 2003; Rudd and Davies, 2002; Visscher and Coe, 2002). McCartney and Rosenthal (2000) note that 'likewise, when reporting research findings to policymakers, data should be presented in a useful and understandable format that addresses their policy concerns' and go on to note that data is seldom analysed in ways that are most useful to policymakers, who are often influenced by compelling argument alone. They later remark that 'policymakers have also turned to social science research to guide their decision making about public expenditures for children's program' (p.172) with, for example, evaluation research informing policymakers about the benefits of programs.

In reporting to the Department for Education and Skills (DfES) the findings from the Effective Provision of Pre-school Education (EPPE) project (Sylva *et al.*, 1999) relating to children's cognitive progress and social behavioural development over the pre-school period (Sammons *et al.*, 2002; Sammons *et al.*, 2003), it was clear that a comparison between the magnitude of the impact of different predictor measures was of particular interest for policy purposes. Therefore, within this context, we started to explore the issue of effect sizes within educational research and, more specifically, school effectiveness research (SER) and complex methodologies such as multilevel modelling.

2.2 Background to the use of effect sizes

The notion of reporting effect sizes is not a new concept. Indeed, a wide range of different types of indices that have generically come to be called 'effect sizes' have existed for a number of years and are most commonly used in experimental studies where there is a control group and an experimental group. Thompson (2002a) categorises the numerous effect size choices into two major classes, namely standardised differences when the relation is assessed via comparison of group means and variance-accounted-for indices when the relation is assessed via the use of correlational approaches.

Standardised differences effect sizes essentially measure the difference in group means divided by some estimate of the standard deviation. This estimate of the standard deviation can be, for example, the 'pooled' standard deviation, as suggested by Cohen (1969). He argues that the standard deviation derived from a larger sample size (i.e. using both the experimental and control group) is a more stable estimate of the population standard deviation. On the other hand, Glass (1976) considered that the standard deviation of the control group was the best estimate of the population mean, claiming that the intervention may have affected the mean and the standard deviation whereas this would not be the case for the control group. For a demonstration of standardised differences effect sizes calculated using different 'standardisers', i.e. different estimates of the standard deviation, see Olejnik and Algina (2000).

Effect sizes linked to variance-accounted-for differences can also be employed although generally less frequently than those related to standardised mean differences. Due to the correlational nature of statistical analyses, a variance-accounted-for relationship effect size similar to r^2 can be calculated to provide an index of the strength of a relationship (see Rosenthal (1994) and Thompson (2002a) for further details). Cohen and Cohen (1983) claim that:

> *one of the most attractive features of MRC (Multiple Regression / Correlation) is its automatic provision of regression coefficients, proportion of variance, and correlation measures of various kinds. These are measures of 'effect size', of the magnitude of the phenomena being studied.*
>
> (Cohen and Cohen, 1983, pp. 6–7)

There has been an active campaign amongst methodologists and applied researchers encouraging authors of academic papers to not just address the question of whether there is a (significant) relationship between two variables but also to take into account the strength of the association and relate this to the practical importance of the findings. Cohen and Cohen (1983) argue that 'the level of consciousness in many areas of just how big things are is at a surprisingly low level. This is because concern about the statistical significance of effects (whether they exist at all) has tended to pre-empt attention to their magnitude' (pp. 6–7). It is important to note that there is a continuing debate relating to the importance attributed to statistical significance tests within research although such a discussion on null hypothesis significance testing (NHST) is outside the scope of this paper. However for further details on the long history of statistical significance tests see, for example, Thompson (2002a) and Fidler (2002) for references relating to the common misuses of NHST.

In 1994 the American Psychological Association (APA) encouraged effect size reporting in the fourth edition of the *APA Publication Manual* (APA, 1994). Despite such encouragement, this advice appeared to have little impact. Therefore, drawing on recommendations for improving statistical practices made by the Task Force on Statistical Inference (TFSI), the 2001 *APA Publication Manual* (APA, 2001) clearly specified a stronger recommendation to report effect sizes:

> *It is almost always necessary to include some index of effect size or strength of relationship in your Results section ... and 'failure to report effect sizes as a defect in the design and reporting of research'*
>
> (APA, 2001, p. 25).

It appears that the APA recommendations have not been fully taken on board by researchers, academics, journal editors, etc. as effect sizes are clearly not being routinely reported in research. This has been a great disappointment to many advocates of statistical reform, particularly because the APA Publication Manual is seen as hugely influential in both setting the standards of editorial practice and as a key step in statistical reform and re-education within psychology (Fidler, 2002). It could be argued though that this scenario has evolved not due to indifference to the concept of effect sizes, rather to a lack of knowledge in the field.

> *Researchers do not know enough about how to compute and report effect sizes ... neither experienced researchers nor experienced statisticians have a good intuitive feel for the practical meaning of common effect size estimates.*
>
> (McCartney and Rosenthal, 2000, p. 176)

Thompson (2002a) suggests that 'progress has been slow also because, until recently, effect sizes computations were not widely available within statistical packages' (p. 67) while Coe (2002a) highlights a general lack of training provided on effect size calculation in standard research methods courses and the non inclusion of effect size formulae in most statistics text books (although Olejnik and Algina (2000) reference a number of American text books on statistical methods that include procedures for computing effect size indices). 'For these reasons, 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' (Coe, 2002a, p. 2).

2.3 The reporting of results in school effectiveness research (SER)

Historically, researchers in the SER field have tended to present their results without reference to the calculation of effect sizes. This may be due to the fact that SER studies almost exclusively sought to model naturally occurring variation in pupil outcomes rather than quantify the impact of a specific intervention. In general, SER has traditionally concentrated far more on measures of school or classroom 'effectiveness', identifying outlier institutions using residual estimates and associated confidence limits. These value-added measures indicate whether pupils' relative progress in different institutions is significantly better or poorer than expected (after control for intake).

In addition, interest has also focussed on statistics showing variance accounted for differences. For example, the intra-school correlation measuring the extent to which the attainment scores of children in the same school resemble each other as compared with those from children at different schools. The reduction in total variance (the proportion statistically 'explained') by a model has also been used, in particular to demonstrate how far the extent of apparent differences between schools in pupil attainment outcomes are accounted for by information about pupil intake (especially prior attainment but also other characteristics such as pupil gender, socio-economic status, ethnicity/language, percentage of pupils eligible for free school meals in a school, etc.). As reported in the Improving Schools Effectiveness Project findings referring to a 6–7 per cent of remaining variance attributable to the schools, 'in percentage terms, this sounds relatively modest but its impact can be of great significance' (Thomas *et al.*, 2001, p. 68). Sammons and Smees (1998) and De Fraine *et al.* (2003) illustrate the way such variance accounted for statistics have been reported in SER studies.

It is of interest to note that in terms of reporting school effects, rather than using the percentage of variance accounted for at the school level, a number of other methods have been suggested but to our knowledge, not widely embraced by the SER field. For example, Jencks *et al.* (1972) introduced the idea of a standardised difference effect size, calculating the difference between the experimental condition and the control group relative to the standard deviation of the criterion variable in the control group condition and then used the square root of the variance accounted for by schools (as quoted by Scheerens and Bosker (1997)). Bosker and Scheerens (1989) suggested that a more interpretable effect standard may be in terms of intervals on the scale of the output variable:

> *Since school effectiveness studies are non-experimental, schools could only be grouped on an* ad hoc *basis in, for example, the highest scoring 20%, the 'middle 60' and the lowest scoring 20%. Though this procedure would inevitably imply exploiting chance, it might still be adopted to make the results of school effectiveness studies amenable to the interpretation of effect sizes according to established convention.*
>
> (Bosker and Scheerens, 1989, p. 247)

The National Institute of Child Health & Development study (NICHD, 2002),[1] evaluated the magnitude of statistically significant childcare effects from a multivariate linear regression model that tested if child functioning at 4 ½ years varied as a function of child-care, quantity, quality and type. The study computed the difference between the adjusted means for high and low group divided by the pooled standard deviation (with continuous variables transformed to categorical ones to obtain high/low groups).

In studies using well-known outcome measures such as GCSE results for which practitioners, researchers and policy makers have an intuitive feel, simple reporting of differences in terms of GCSE grade differences has often been thought to be sufficient. An example of this is the Forging Links research (Sammons *et al.*, 1997), which explored differential school effectiveness in a sample of inner London secondary schools, controlling for prior attainment at age 11, with GCSE results at age 16 years as outcomes for three consecutive cohorts. The difference between the most and the least effective school (in terms of value added residual estimates) was relatively large at 12 GCSE score points. This was reported for an average student as equivalent to the difference between gaining six GCSE Grades Bs rather than six Grade Ds. It was also reported that a sizeable proportion of schools (11 out of 69) had significantly different results equivalent to +/– 10 GCSE points. The multilevel fixed effects estimates of background factors controlled for in the model were also presented. These are estimates of the amount by which the outcome (in this case GCSE points) changes, on average, relative to one unit of change in the background variable when all other measures in the model are controlled.

For example, the impact of the low income indicator (eligibility for free school meals (FSM) versus not eligible) was –3.3 points indicating that, on average, pupils eligible for FSM achieved approximately 3 GCSE points less than pupils not eligible for FSM. This could be described as the difference between three Grade Cs rather than three Grade Ds. For gender, there was on average a 2.1 GCSE points difference between girls and boys (in favour of girls). Therefore, it was concluded that the net impact of eligibility for FSM on GCSE total point score was greater than the size of the gender gap. The relative importance of gender and FSM was directly comparable because the outcome measure (total GCSE point score) could be interpreted in a meaningful way.

Although fixed effects estimates can be interpreted in the ways outlined above when well-known outcomes are studied, the values can be difficult to compare due to the different units involved. In reporting their 'Playing for Success' evaluation findings, Sharp *et al.*, (2003) converted fixed effects estimates into normalised coefficients which represent the correlation between each variable and the outcome taking account of the other variables in the model. In other words, these normalised coefficients indicate the 'strength' of each relationship, allowing the different predictors to be compared in terms of their influence on the outcome, when all other predictors are simultaneously taken into account. The NICHD study noted above also calculated a complementary measure of association with structural coefficients, reflecting 'the relative predictive power of each predictor included in the analysis model without adjusting for shared variance among the predictors' (NICHD, 2002, p. 150).

A further difficulty arises when a range of different and less well-known outcome measures are employed. It is here that the use of effect sizes has attractions by offering the possibility of a readily interpretable universal indicator that can enable comparisons across different studies involving a range of outcome measures (see Coe, 2002a for a clear discussion of this point). By indicating the relative importance of different measures in such a way, the research knowledge base may be improved and, in turn, research may become more accessible to other stakeholders such as policy makers. However, it may also be that the very nature of some complex statistical analyses which involve many predictors, lead researchers to be cautious about identifying an effect size, recognising that effect sizes will often depend greatly on the particular model specification used.

2.4 Calculating effect sizes within multilevel modelling

The EPPE study was commissioned and funded by the Department for Education and Employment (DfEE) now the Department for Education and Skills (DfES). Hence the research team report findings direct to policy makers. Stimulated by questions raised relating to children's cognitive progress and social behavioural development over the pre-school period (Sammons *et al.*, 2002; Sammons *et al.*, 2003; Sammons *et al.*, forthcoming, a), we sought to include an effect size statistic along with other information such as percentage of total variance explained, intra-school correlation residual estimates of pre-school centre effectiveness, etc. As the EPPE study is not an experimental study, rather it explores naturally occurring variation in pre-school provision, an educational effectiveness design was adopted. This approach sought to ensure proper control for the influence of intake differences and testing of process measures of interest (type of provision attended, quality indicators, duration of pre-school experience, etc.).

Multilevel models were employed to separate the pre-school centre level variance in child outcome measures from that attributable to differences at the individual child level, recognising the hierarchical nature of the data (Goldstein, 1995). Of particular interest was the calculation of pre-school centre-level residuals for a range of outcomes, which were used to assist in the selection of pre-school centres for in-depth case study (Siraj-Blatchford *et al.*, 2003). The EPPE study thus adopted a mixed methods research design linking qualitative and quantitative methodologies to illuminate the study of pre-school processes and pedagogy (for further discussion of methodological aspects, see Sammons *et al.*, forthcoming, b).

In multilevel modelling, the 'fixed' part of the model (i.e. the fixed effect estimate) is essentially the mean difference between two groups, after statistically controlling for the influence of other factors. In other words, the estimates show the mean difference, net of the impact of other explanatory measures specified in the model.[2] Thus, since 'standardised differences effect sizes' fundamentally measure the difference in group means divided by some estimate of the outcome (dependent variable) standard deviation, the fixed effect estimate derived from the multilevel analysis can be divided by some estimate of the standard deviation to calculate an effect size. The dilemma is which standard deviation: raw or residual (adjusted) standard deviation. Employing the raw outcome standard deviation (i.e. amount of variation in the outcome measure before appropriate controls have been made) links to the standardised regression coefficient formula[3] often calculated in multiple regression analyses.

In multilevel analyses, residuals are calculated at each level of the model, thus the number of possible residual standard deviations depends on the number of levels specified. Say that pupils are specified at level 1 (as is the case in many educational effectiveness models), the pupil level variance is the amount of variation in the outcome measure attributable to the individual pupil after appropriate controls have been made. It is important to note that using the level 1 residual standard deviation tends to *increase* the effect size compared with calculations that employ a raw standard deviation. However, such calculations are considered appropriate because they explicitly model the extent and impact of clustering in the data.

After discussion and comparison of different approaches, for categorical dependent variables,[4] the effect sizes reported in EPPE were calculated following the method outlined by Tymms *et al.* (1997) in their investigation of the attainment and progress of pupils in the first year of school. Strand (2002) also used the same method when reporting the strength of association between pupil mobility, attainment and progress during key stage 1. The advantage of this approach is that it directly employs the fixed effect estimates (i.e. predictor coefficients) based on the multilevel analyses which takes account of the hierarchical structure of the data.

Effect size (ES) = categorical predictor variable coefficient / √child level variance

or $$\Delta = \frac{\beta_1}{\sigma_e}$$

or *the difference between the estimated means for the groups defined by the dummy codings 1 and 0 expressed as a fraction of the pupil level standard deviation, after appropriate controls have been made* (Tymms *et al.*, 1997, p. 112).

In order to obtain continuous predictor variable effect sizes (i.e. coefficients from a multilevel model expressed in standard deviations as scale units), Snijders and Bosker (1999) calculated standardised coefficients following the standardised regression coefficient formulae from multiple regression:

ES = continuous predictor variable coefficient*SD continuous predictor variable / SD dependent variable

Our effect size calculation for continuous predictor variables has been based on the basic principle adopted by Snijders and Bosker (1999) but the raw standard deviation has been replaced by the level 1 residual standard deviation. This ensures consistency in approach between the effect size formulae for categorical and continuous predictor variables. In addition, as the predictor variables have not been normalised (and thus the standard deviation is not necessarily equal to 1), the formula recommended (Sammons *et al.*, 2002; Sammons *et al.*, 2003) has been revised to show the standard deviation multiplied by 2.

ES = continuous predictor variable coefficient*2SD continuous predictor variable / √child level variance

or $$\Delta = \frac{\beta_1 * 2sd_{x1}}{\sigma_e}$$ where $_{x1}$=continuous predictor variable

This effect size describes the change on the outcome measure that will be produced by a change of +/– one standard deviation on the continuous predictor variable, standardised by the within school standard deviation adjusted for covariates in the model.

These formulae outlined above for the calculation of effect sizes for categorical and continuous predictor variables employed in a multilevel analyses have the advantage of being relatively quick to calculate and readily understandable.

2.5 The presentation and interpretation of effect sizes

The multi-representation of results is generally considered to aid interpretation. Therefore, it is recommended that effect sizes are presented in tabular, graphical and textual format. Note though that separate charts for effect sizes relating to categorical and continuous predictor variables should be produced because the different calculations result in the two types of effect sizes that are not directly comparable. In terms of graphs relating to effect sizes for categorical variables, the number of units in each category (in SER studies, generally children) should be shown as effect sizes for some categories may only apply to a small number of units. Whether an effect for a predictor variable is positive or negative is also of great importance and should therefore be specified.

As with the reporting of value-added residual estimates, the use of confidence intervals indicating the statistical uncertainty attached to any effect size would greatly aid interpretation. However, as Thompson (2002a) comments the reporting of confidence intervals for effect sizes is ‘an appealing strategy, but estimating these intervals can be very complicated’ (p. 69). Coe (2002a) also argues for the calculation of effect sizes stating that:

> *confidence intervals generally convey the same information as the more widely used tests of statistical significance, but avoid the need for a usually inappropriate yes/no decision about whether there is an effect, instead allowing that effect to be quantified within a given margin of error.*
>
> (Coe, 2002a, p. 8)

This also raises the issue of inclusion of non-significant variables in terms of model specification as there may be cases where a non-statistically significant variable may still improve model fit and is shown to have an important practical impact on the outcome measure. ‘It is important to note that an effect size estimate can be computed regardless of whether “significance” is obtained’ (McCartney and Rosenthal, 2000, p. 175). Therefore, when presenting effect sizes, it is important to display the statistical significance of each predictor variable and, in the case of categorical predictor variables, each dummy variable.

In terms of the interpretation of effect sizes, Cohen (1969) provides a general rule (in particular for *d* or other effect sizes that can be converted to *d*). He suggests that a standardised difference of |0.8| may be considered large, |0.5| medium and |0.2| small. However, it has been argued that these tentative benchmarks need to be considered with considerable caution as there is little empirical justification for these standards (Olejnik and Algina, 2000). Furthermore, in relation to SER studies modelling naturally occurring variation in pupil outcomes, it is worth noting that 'effect sizes in naturalistic studies are typically small because they are measured in the context of many influences' (NICHD, 2002, p. 136) quoting from Cohen (1988). The NICHD study reports that:

> *even modest effects may aggregate when large numbers of children are affected. For example, many of the most important risk behaviours from a public health perspective have a low or moderate risk, but they are multiplied in importance because of their wide prevalence and links to problematic outcome (Jeffrey, 1989).*
>
> (NICHD, 2002, p. 158)

Thompson (2002b) notes that:

> *the overly rigid use of fixed benchmarks for small, medium and large effects fails to consider the possibility that small, replicable effects involving important outcomes can be noteworthy, or that large effects involving trivial outcomes may not be particularly noteworthy.*
>
> (Thompson, 2002b, p. 30)

Furthermore, Gage (1984) agrees with this, defending:

> *the proposition that correlations or differences do not need to be large in order to be important. In education, we are not influencing life or death. But we are influencing dropout rates, literacy, placement in special classes, love of learning, self-esteem and the holistic ability to integrate many facts and concepts in a complex way. The implications of research for practice depend not on the size of the effects but on the costs and benefits of any change in practice.*
>
> (Gage, 1984, p. 90)

Furthermore, Glass, McGaw and Smith (1981) confirm that:

> *in education, if it could be shown that making a small and inexpensive change would raise academic achievement by an effect size of even as little as 0.1, then this could be a very significant improvement,*

> *particularly if the improvement applied uniformly to all students, and even more so if the effect were cumulative over time.*
>
> (Glass, McGaw and Smith, 1981, p. 104)

For a policy perspective on the sometimes beneficial impact of effect sizes within the 'small' category in Cohen's rule of thumb framework, see a summary of comments from DfES colleagues in Chapter 10.

Consideration should also be given to the issue of appropriate levels of control in the model when interpreting the effect size. For further discussion relating to the possible inflation or underestimation of effects when selection factors are not adequately controlled or over controlled, see NICHD (2002). It is widely documented that when considering the size of effects, it is important to take into consideration other factors such as the context of the study, outcome(s) studied and predictors controlled for.

2.6 Conclusion

It has been argued that 'most social scientists seldom analyze data in ways that are most useful to policy makers' (McCartney and Rosenthal, 2000, p. 173). Our exploration of the possibilities and limitations in employing effect size within multilevel analyses was stimulated by requests from a policy audience to present the EPPE findings in a more useful and accessible format. We employed the method outlined in the paper for calculating and displaying effect sizes to evaluate the impact of different influences on child outcomes. As educational researchers, as opposed to statisticians, we are extremely conscious that the calculation of effect sizes within methodologies such as multilevel modelling is a complex area. Chapters 3–5 will detail further the technical aspects of the calculations and alternative approaches.

We tend to agree with Coe (2002b) that a more wide-spread use of effect size measures would probably be advantageous, although Olejnik and Algina (2000) comment that effect sizes are not without their critics who argue whether such measures actually contribute to a better understanding of a study's results. Effect sizes may well prove very useful in certain contexts and can provide an additional indicator in the interpretation of statistical analyses, although they should not be seen as a statistical 'cure all'. It is important that other statistics are referenced in SER studies such as the percentage of total variance accounted, intra-school correlations, fixed effect estimates for a set of predictor measures with their associated standard errors and indicators of statistical significance. From a situation where effect sizes have not been routinely reported, it would be unfortunate if the pendulum swung to the other extreme where effect sizes are considered as the only important or interesting measure of policy or practical relevance in reporting educational research.

In addition, it is vital to recognise that any effect size will only be as good as the model from which it is derived. The set of predictor measures available for analysis will have a major impact on the estimates obtained and therefore the effect sizes calculated. When comparing effect sizes from a relatively simple multilevel model with those from a more detailed complex multilevel model with better controls, differences in effect sizes may emerge which are in fact a direct result of variations in model specification rather than any real differences in impact. For example, studies may show a larger effect size for family socio-economic status (SES) when there is no control for other influences such as mother's level of qualification and the home learning environment. When such predictors are controlled, the size of the family SES effect is often much reduced. Although not referring to analyses using multilevel models, Olejnik and Algina (2000) emphasise this point stating that measures of effect sizes can be affected by the research design used.

It is also important to note that, even within the same study, a given set of predictor measures may show different relationships with different outcomes as certain outcome measures (e.g. reading or language) are more sensitive to certain background influences than others. Such differences can be of both theoretical and practical interest. Furthermore, relationships may change over time as shown by Sammons (1995) who conducted further analyses of the *School Matters* data (Mortimore *et al.*, (1988)) following children up to age 16 years. The same set of predictor measures and outcome tests were employed at different time points and the research showed that it was possible to establish whether, taken together, background influences reduced or stayed the same in terms of impact on attainment at different time points. In addition, changes in the estimates (though not in terms of effect sizes) were reported for specific characteristics. It was found that at age seven years the same set of predictor measures accounted for relatively similar proportions of total variance for reading and mathematics (19.6 per cent and 18 per cent respectively). Just under three years later, however, the same set of predictors accounted for 20.6 per cent of the total variance in the same children's reading scores, but only 11.3 per cent for mathematics. In other words, the impact of background measures reduced significantly at the later time point for mathematics but slightly increased for the reading measure. Thus control for the relevant set of background measures was especially important when interpreting reading differences for the older age group.

A large and appropriate sample, attention to the reliability and validity of the outcome measures used, good control for relevant prior attainment and background measures, and the use of multilevel modelling that capitalises on the hierarchical nature of the data, all remain essential in educational effectiveness research focussing on pupil progress. It should be remembered that relatively modest effect sizes may be of educational significance if they relate to measures amenable to policy influence, whereas larger effect sizes, if they apply to measures that are difficult to alter, may be of less relevance. In all cases, research judgement and additional details will be essential to aid interpretation of findings from such analysis.

To summarise, we suggest that effect size calculations can provide additional useful data for researchers and policy makers when interpreted with caution in conjunction with other important statistical measures and indicators, but should not be seen as offering the only gold standard in reporting educational research. In conclusion, although 'an effect size provides a first step towards evaluating the practical importance of a finding' (McCartney and Rosenthal, 2000, p. 174) it should never be reported in isolation or treated at face value without careful consideration of context (especially the sample), the nature of the outcome and model, background measures employed, formula used to calculate the effect size, and its particular rationale and application for the purpose of the research.

Endnotes

[1] The NICHD study is a prospective longitudinal study of more than 1000 children in America.

[2] The main guide to the inclusion/exclusion of different explanatory measures in a multilevel model is a comparison of the explanatory measure estimate to its standard error as with a sufficiently large random sample the ratio of a fixed parameter to its standard error should be approximately normally distributed with mean 0 and variance 1. Note that for random parameters, the likelihood ratio statistic is often cited as a better test, indicating the 'goodness of fit'

of different models.

[3] The standardised regression coefficients in multiple regression is calculated as the regression coefficient for the predictor variable * sd_{x1} / sd_y.

[4] Note that, in multilevel models, the method of dummying variables is used for a categorical variable. In other words, if there are n groupings within a category with each child only ever assigned to one group, a base group is identified and the remaining n-1 groups are defined as dummy variables. The value of the dummy variable is equal to 1 if the child belongs to the group or 0 otherwise. Those children in the base group have a 0 assigned for all n-1 dummy variables.

References

AMERICAN PSYCHOLOGICAL ASSOCIATION (1994). *Publication Manual of the American Psychological Association.* Fourth edn. Washington, DC: Autor.

AMERICAN PSYCHOLOGICAL ASSOCIATION (2001). *Publication Manual of the American Psychological Association.* Fifth edn. Washington, DC: Autor.

BOSKER, R. and SCHEERENS, J. (1989). 'Criterion-definition, effect size and stability, three fundamental questions in school effectiveness research.' Paper presented at the International Congress for School Effectiveness and School Improvement, Rotterdam.

COE, R. (2002a). 'What is an effect size?' *Journal of the Economic & Social Research Council Teaching and Learning Programme Research Capacity Building Network*, 6–8.

COE, R. (2002b). 'It's the effect size, stupid: what effect size is and why it is important.' Paper presented at the British Educational Research Association Annual Conference, Exeter, 12–14 September [online]. Available: http://www.leeds.ac.uk/educol/documents/00002182.htm
[4 March, 2004].

COHEN, J. (1969). *Statistical Power Analysis for the Behavioural*

Sciences. London: Academic Press.

COHEN, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York: Laurence Erlbaum.

COHEN, J. and COHEN, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioural Sciences*. Second edn. Mahwah, NJ: Lawrence Erlbaum Associates.

DE FRAINE, B., VAN DAMME, J., VAN LANDEGHAM, G., OPDENAKKER, M. and ONGHENA, P. (2003). 'The effect of schools and classes on language achievement', *British Educational Research Journal*, **29**, 6, 841–59.

DEMIE, F. (2003). 'Using value-added data for school self-evaluation: a case study of practice in inner city schools', *School Leadership & Management*, **23**, 4, 445–67.

DUDLEY, P. (1999). 'Primary schools and pupil "data".' In: SOUTHWORTH, G. and LINCOLN, P. (Eds) *Supporting Improving Primary Schools: the Role of Heads and LEAs in Raising Standards*. London: Falmer Press.

ELLIOT, K. and SAMMONS, P. (2001). 'Using pupil performance data: three steps to heaven?' *Improving Schools*, **4**, 54–65.

ELLIOT, K. and SAMMONS, P. (2003). 'From data rich to information rich', *Professional Development Today*, **7**, 14–18.

FIDLER, F. (2002). 'The fifth edition of the APA publication manual: why its statistical recommendations are so controversial', *Educational and Psychological Measurement*, **62**, 5, 749–70.

GAGE, N.L. (1984).'What do we know about teaching effectiveness?' *Phi Delta Kappan*, **66**, 87–93.

GLASS, G.V. (1976). 'Primary, secondary, and meta-analysis of research', *Educational Researcher*, **5**, 10, 3–8.

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

GOLDSTEIN, H. (1995). *Multilevel Statistical Models*. Second edn. London: Arnold.

JEFFREY, L. (1989). 'Risk behaviours and health. Contrasting individual and population perspectives', *American Psychologist*, **44**, 1194–202.

JENCKS, C., SMITH, M., ACLAND, H., BANE, M. J., COHEN, D., GINTIS, H., HEYNS, B. and MICHELSON, S. (1972). *Inequality: a Reassessment of the Effect of Family and Schooling in America*. London: Penguin Books.

KLUGER, A. and DeNISI, A. (1996). 'The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory', *Psychological Bulletin*, **119**, 254–84.

McCARTNEY, K. and ROSENTHAL, R. (2000). 'Effect size, practical importance, and social policy for children', *Child Development*, **71**, 173–80.

MORTIMORE, P., SAMMONS, P., STOLL, L., LEWIS, D. and ECOB, R. (1988). *School Matters: the Junior Years*. Wells: Open Books.

NATIONAL INSTITUTE OF CHILD HEALTH AND DEVELOPMENT (2002). 'Early child care and children's development prior to school entry: results from the NICHD study of early child care', *American Educational Research Journal*, **39**, 133–64.

OLEJNIK, S. and ALGINA, J. (2000). 'Measures of effect size for comparative studies: applications, interpretations, and limitations', *Contemporary Educational Psychology*, **25**, 241–86.

ROSENTHAL, R. (1994). 'Parametric measures of effect size.' In: COOPER, H. and HEDGES, L.V. (Eds) *The Handbook of Research Synthesis*. New York, NY: Russell Sage Foundation.

RUDD, P. and DAVIES, D. (2002). *A Revolution in the Use of Data? The LEA Role in Data Collection, Analysis and Use and Its Impact on Pupil Performance* (LGA Research Report 29). Slough: NFER.

SAMMONS, P. (1995). 'Gender, ethnic and socio-economic differences in attainment and progress: a longitudinal analysis of student achievement over nine years', *British Educational Research Journal*, **21**, 4, 465–85.

SAMMONS, P., ELLIOT, K., SYLVA, K., MELHUISH, T., SIRAJ-BLATCHFORD, I., TAGGART, B. and SMEES, R. (forthcoming, a). 'The impact of pre-school on young children's cognitive attainment at entry to reception' (Special Issue: Early Years), *British Educational*

Research Journal.

SAMMONS, P., SIRAJ-BLATCHFORD, I., SYLVA, K., MELHUISH, E., TAGGART, B. and ELLIOT, K. (forthcoming, b). 'Investigating the effects of pre-school provision: using mixed methods in the EPPE research' (Special Issue: Theory and Practice), *International Journal of Research Methods.*

SAMMONS, P. and SMEES, R. (1998). 'Measuring pupil progress at key stage 1: using baseline assessment to investigate value added', *School Leadership & Management*, **18**, 3, 389–407.

SAMMONS, P., SYLVA, K., MELHUISH, E. C., SIRAJ-BLATCHFORD, I., TAGGART, B. and ELLIOT, K. (2002). *The Effective Provision of Pre-School Education (EPPE) Project: Technical Paper 8a – Measuring The Impact of Pre-School on Children's Cognitive Progress Over the Pre-School Period.* London: DfES and University of London, Institute of Education.

SAMMONS, P., SYLVA, K., MELHUISH, E. C., SIRAJ-BLATCHFORD, I., TAGGART, B. and ELLIOT, K. (2003). *The Effective Provision of Pre-School Education (EPPE) Project: Technical Paper 8b – Measuring the Impact of Pre-school on Children's Social/Behavioral Development Over the Pre-School Period.* London: DfES and University of London, Institute of Education.

SAMMONS, P., THOMAS, S. and MORTIMORE, P. (1997). *Forging Links: Effective Schools and Effective Departments.* London: Paul Chapman.

SAUNDERS, L. (2000). 'Understanding schools' use of "value added" data: the psychology and sociology of numbers', *Research Papers in Education*, **15**, 3, 241–58.

SCHEERENS, J. and BOSKER, R. (1997). *The Foundations of Educational Effectiveness.* Oxford: Pergamon.

SHARP, C., KENDALL, L. and SCHAGEN, I. (2003). 'Different for girls? An exploration of the impact of Playing for Success', *Educational Research*, **45**, 3, 309–24.

SIRAJ-BLATCHFORD, I., SYLVA, K., TAGGART, B., SAMMONS, P., MELHUISH, E. and ELLIOT, K. (2003). *The Effective Provision of Pre-School Education (EPPE) Project: Technical Paper 10 – Intensive Case Studies of Practice Across the Foundation Stage.* London: DfES and

University of London, Institute of Education.

SNIJDERS, T. and BOSKER, R. (1999). *Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.

STRAND, S. (2002). 'Pupil mobility, attainment and progress during key stage 1: a study in cautious interpretation', *British Educational Research Journal*, **28**, 1, 63–78.

SYLVA, K., SAMMONS, P., MELHUISH, E. C., SIRAJ-*BLATCHFORD, I. and TAGGART, B. (1999). The Effective Provision of Pre-School Education (EPPE) Project: Technical Paper 1 – an Introduction of EPPE*. London: DfEE and University of London, Institute of Education.

THOMAS, S., SMEES, R., SAMMONS, P. and MORTIMORE, P. (2001). 'Attainment, progress and added value.' In: MacBEATH, J. and MORTIMORE, P. (Eds) *Improving School Effectiveness*.

THOMPSON, B. (2002a). '"Statistical," "practical" and "clinical": how many kinds of significance do counselors need to consider?' *Journal of Counselling and Development*, **80**, 64–71.

THOMPSON, B. (2002b). 'What future quantitative social science research could look like: confidence intervals for effect sizes', *Educational Researcher*, **31**, 325–32.

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.

VISSCHER, A. and COE, R. (Eds) (2002). *School Improvement through Performance Feedback*. Lisse: Swets & Zeitlinger.

YANG, M., GOLDSTEIN, H., RATH, T. and HILL, N. (1999). 'The use of assessment data for school improvement purposes', *Oxford Review of Education*, **25**, 4, 469–83.

3 Presenting the results of complex models – normalised coefficients, star wars plots and other ideas

Ian Schagen

3.1 Introduction

At the end of the twentieth century the introduction of multilevel modelling techniques and access to large and complex datasets gave statisticians working in educational research power to carry out complex and sophisticated analyses, which enabled us to gain important insights into many educational processes (see Schagen & Hutchison, 2003). However, in many ways our ability to present and explain our results has not kept pace with our power to carry out analyses and the obscurity of many presentations has led to a backlash. There is a movement, with some adherents even among respected academics, which seems to be saying: 'No-one can understand the results of multilevel analysis, so it's not worth doing' (see e.g. Jesson, 2003; Gorard, 2003). It is possible to dub this attitude 'Keep it simple even if it's wrong', but being rude about it is not helpful and not the way to convince people that sophisticated analyses are really worth doing.

An example which shows that simple analysis gives misleading and incorrect results and that it takes complex and sophisticated modelling to uncover the true relationship, is given by the debate on the effect of class size on performance. It is frequently the case that simple analysis of pupils' test scores shows that those in larger classes get higher scores, on average. A previous Chief Inspector of Schools immediately used this kind of result to argue that it was pointless spending money on reducing class sizes, as the data showed that pupils do better in bigger classes. Of course, things are not as simple as that – higher attaining pupils tend to be grouped into bigger classes, while their lower attaining colleagues may get special treatment in smaller groups. Only more recently, with sophisticated analysis, have researchers managed to demonstrate that taking careful account of such effects shows that smaller classes can improve pupils' scores (see Blatchford *et al.*, 2002).

The onus is really on the educational researchers and statisticians who believe in the power of multilevel analysis to derive ways of presenting its results which are valid and also accessible to intelligent but non-technical people within the whole field of education. Gorard (2003, p. 54) is right when he complains that too often multilevel modelling results are published as indigestible tables of coefficients with obscure variable names attached and with no explanation of what the coefficients mean or of what the most important factors are in terms of their impact on the outcome of interest.

At NFER we have been involved in multilevel analysis on a routine basis for a number of years and have evolved several ways of presenting results. These are not perfect and it is time that they were exposed to a critical audience so that they can be refined and improved (or replaced). They also raise a number of issues that should be debated within a wider forum.

In this paper I will first discuss some possible ways in which multilevel model outcomes may be presented, based on examples of analyses which have been carried out.

3.2 Presenting multilevel outcomes for the fixed part of the model

The fixed part of a multilevel model is often, though not exclusively, the part which is of most interest to a general audience. It represents the overall or average differences in outcomes which may be attributed to each background variable, controlling for all the rest. The model gives us directly the first and simplest way of presenting these impacts – their coefficients.

3.2.1 Coefficients

The coefficient (B let us call it) relating background measure X to outcome Y is the average change in Y associated with one unit change in X, taking account of all the other variables in the model. If X is a binary measure (say 0 for boys and 1 for girls), then B is relatively easy to interpret as the average difference in outcome Y between girls and boys (as ever, controlling for other variables).

Two problems arise with this. The first is related to the units of Y, which may obscure the educational significance of the value of B. If these are, say, total score points, then how do we know whether a value of B of 1.2 is important or not? We will return to this issue later. The second problem is concerned with comparing the relative coefficients of different variables.

Suppose B is 1.2 for the girl/boy effect and 0.12 for the effect of percentage known to be eligible for free school meals (FSM), then which of these factors is the more important in terms of its impact on Y? The first is 10 times the second and so seems on the surface to relate to the larger educational impact. But the coefficient of FSM is the average change in Y per unit change in FSM. The girl/boy variable is either 0 or 1, but the FSM variable may range from 0 to 80, hence it seems possible that its overall impact across the range might well be larger. To try to deal with this issue we develop a new measure – the normalised coefficient.

3.2.2 Normalised coefficients

To compare the overall effects of different variables, we need measures which are dimensionless. One such measure is obtained by scaling the coefficients by the standard deviations of both X and Y:

$$n = 100*B*s/S, \qquad (1)$$

where s = standard deviation of X
S = standard deviation of Y.

The 'normalised coefficient' n represents the expected change in Y (expressed as a percentage of the standard deviation in Y) for one standard deviation change in X. It has no units and can be compared directly with the normalised coefficients for other variables. Table 3.1 shows the computation of these values for four different variables, based on analysis of national data linking pupils' performance at KS1 in 1998 to their KS2 results in 2002 (see Schagen & Benton, 2003). The outcome variable is average KS2 point score (standard deviation = 4.74).

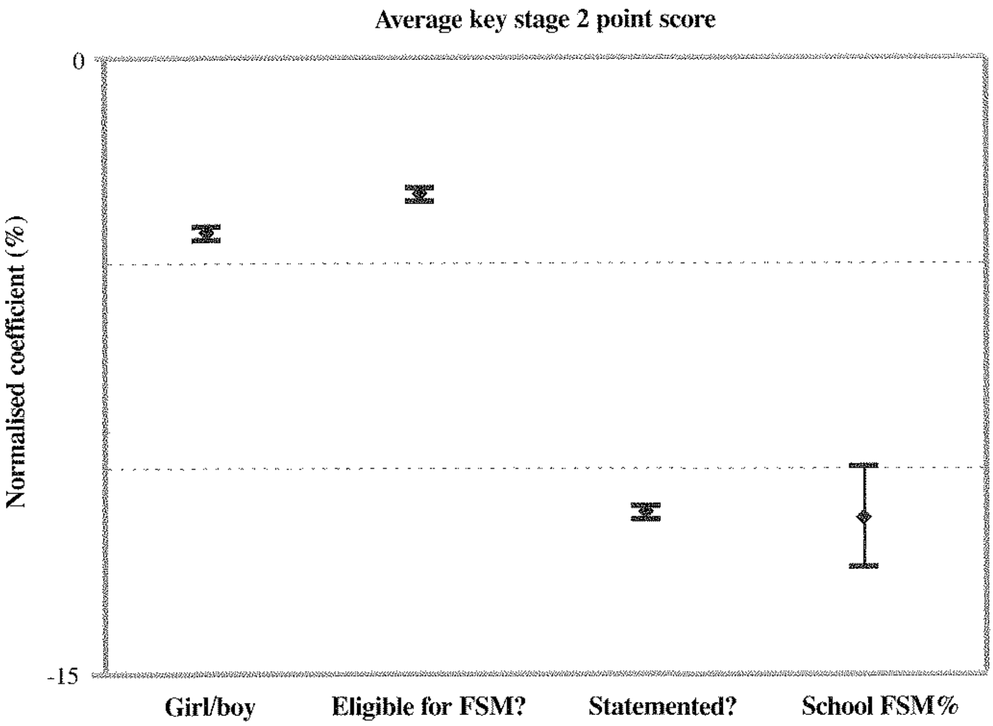
The values in the final column give us a means of comparing the overall impact of the different variables on the outcome; it seems that school-level FSM has a larger impact than the pupil-level variable, but that the effect of statemented pupils is the largest overall.

Table 3.1 Normalised coefficients for selected background variables related to KS2 average point score outcome in 2002

Description	Standard deviation	Range	Coefficient	Standard error of coefficient	Normalised coefficient
Girl/boy	0.5	0,1	-0.411	0.007	-4.33
Eligible for FSM?	0.37	0,1	-0.436	0.010	-3.41
Statemented?	0.15	0,1	-3.593	0.030	-11.37
School FSM %	14.38	0,100	-0.036	0.002	-10.81

Figure 3.1 below is a way of displaying these normalised coefficients and also includes the 95% confidence interval for each.

Figure 3.1 Graphical example of normalised coefficients



These plots can be quite good ways of seeing the relative impacts of a range of background variables, and are sometimes informally known as 'star wars plots' (look at the diagram sideways to see why). Any variable whose coefficient is not significant at the 5% level will have a 95% confidence interval which straddles the zero axis on this plot. I believe that the inclusion of confidence intervals in the presentation of results is vitally important to give our audiences a full understanding of the uncertainty in our models.

3.2.3 Pseudo effect sizes

One way of conceptualising these 'normalised coefficients' is as representations of the overall impact, across the whole population, of each factor: sex, individual FSM, statemented pupils, and school-level FSM. But the overall impact of statemented pupils is relatively small partly because there are few of them; what about the impact of this factor considered in terms of the expected change brought about by being statemented rather than not statemented? For binary variables it is straightforward to compute a 'pseudo-effect size':

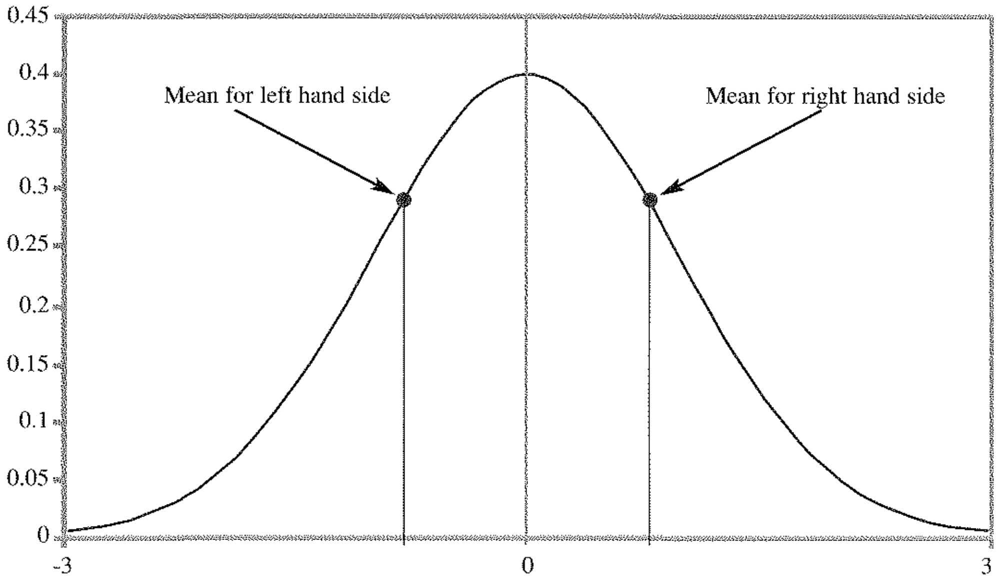
$$e \quad = \quad 100{*}B/S \qquad (2)$$

The index *e* therefore shows the impact of going from the 'low' to 'high' value of the binary variable, as a percentage of the standard deviation in the outcome variable. This gives a simple way of comparing binary variables, but can we find an equivalent for the others? We could just use *e* for binary and n for the rest, but are we sure we would be comparing like with like? For binary variables we are thinking about the impact of switching between two states; non-binary variables have more states to switch among, so what might we think of as the 'average switch' for them?

It is not immediately clear how we should approach this – there are a number of ways, of which we shall briefly consider three. Let us begin by considering a variable X with a Standard Normal distribution (mean 0, standard deviation 1), as in Figure 3.2. Suppose now we 'discretise' it, i.e. convert it into a binary variable with all values less than zero having a value 0 and all those above zero having a value 1. The mean values for these two parts of the distribution are shown in the figure, and have values of approximately -0.79 and 0.79 respectively.

The distance between these two means is therefore 1.58 and thus it might be reasonable to use this as the factor to convert the effect of a continuous variable into the equivalent for a binary variable. Alternatively, we could consider the medians of the two halves of the distribution, i.e. the first and third quartiles of the distribution. These are at -0.675 and 0.675 respectively, giving a total distance of 1.35.

Figure 3.2 Normal Distribution discretised into two parts



Finally, we might consider a slightly different approach. Suppose we randomly pick two separate cases with different values of X – what do we expect the difference between them to be? The expected value of the absolute difference between two values is not entirely straightforward to compute, but a simple alternative is the 'root mean square' expected difference – the square root of the expected value of the square of the difference. If X has standard deviation equal to 1, then the expected value of the square of the difference is 2 and the square root is √2, which equals 1.41.

From the above, there are three different values for the factor we could apply to the effect size for non-binary variables to compare with that for a binary variable:

- 1.58 (based on distance between 'split means')
- 1.35 (based on distance between 'split medians')
- 1.41 (based on root mean square expected difference between two random values).

The final value is in between the other two, so for this reason we shall use it to derive the formula for pseudo-effect size for non-binary variables, which now becomes:

$$e \quad = \quad 100{*}B{*}\sqrt{2}s\,/S, \qquad (3)$$

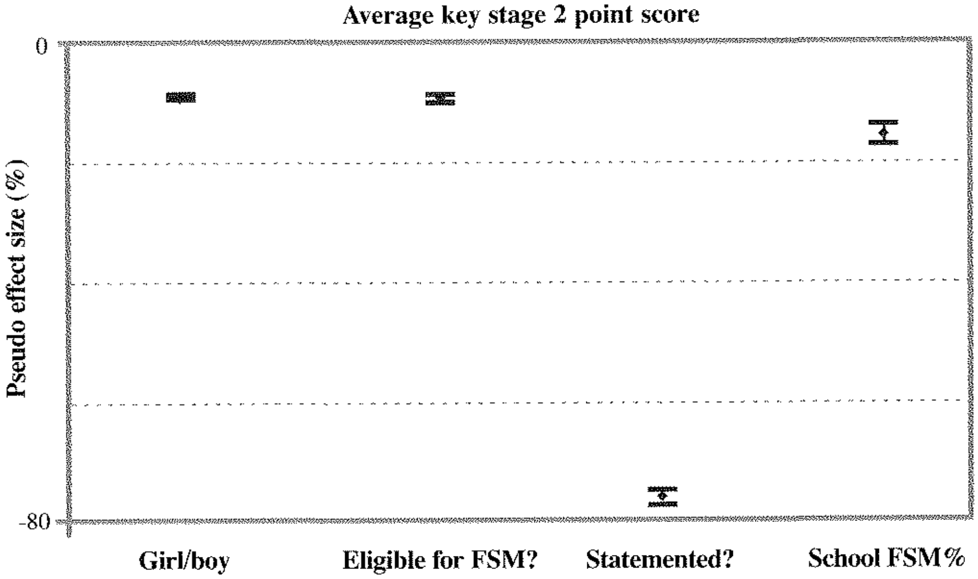
Table 3.2 shows the values of pseudo-effect size for the four variables we have been considering, and these are plotted in Figure 3.3.

Table 3.2 Pseudo-effect sizes for selected background variables related to KS2 average point score outcome in 2002

Description	Standard deviation	Range	Coefficient	Standard error of coefficient	Pseudo effect size
Girl/boy	0.5	0,1	-0.411	0.007	-8.67
Eligible for FSM?	0.37	0,1	-0.436	0.010	-9.21
Statemented?	0.15	0,1	-3.593	0.030	-75.80
School FSM %	14.38	0,100	-0.036	0.002	-15.28

The main difference to be observed (apart from a general change of scale) is that the impact of being statemented has greatly increased relative to the other factors. This reflects the fact that being statemented has a big effect on those pupils, even if its overall effect is more restricted.

Figure 3.3 Graphical example of pseudo-effect sizes



3.2.4 Measures for interaction terms

The measures we have considered so far relate to what might be termed 'main effects', that is the direct impact of background factors on the outcome variable. In many modelling situations we also include interaction terms, which look at the relationship between two (or possibly more) variables considered together. For example, we may be interested in whether boys and girls have different relationships with prior attainment and thus construct an interaction term in the model which relates to this. Let us consider two interaction terms from the KS1 to KS2 model we are considering:

SEXINT Interaction between boy/girl (binary) factor and average KS1 score

FSMINT Interaction between school FSM % and average KS1 score

We could treat these in the same way as the other variables, but it is not so clear how to interpret the measures. It seems sensible to define one of the interacting variables as in some sense 'more basic' than the other and to look at the interaction coefficient relative to the main coefficient of this variable. In the above examples, it would seem that prior attainment is the more basic, and we are interested in how the other variables influence the slope of the line relating prior attainment to the KS2 outcome. We therefore suggest that the 'interaction pseudo effect size' be computed as:

$$\begin{aligned} I &= 100^*B/K && \text{if the other variable is binary} \qquad (4) \\ &= 100^*B^*\sqrt{2}s/K && \text{if the other variable is not binary} \end{aligned}$$

Where B is the coefficient of the interaction term,
K is the main coefficient of the more basic variable,
s is the standard deviation of the other variable.

In our example, the more basic variable is prior attainment, measured in terms of KS1 average score, and the main coefficient is 0.856 (computed by summing coefficients of its components). Table 3.3 and Figure 3.4 show the results for these two example interaction terms.

Table 3.3 Pseudo-effect sizes for interaction terms with key stage 1 average score, related to key stage 2 average point score outcome in 2002

Description	Standard deviation of other variable	Coefficient	Standard error of coefficient	Pseudo-effect size
Boy/girl v. KS1 average	0.5	0.0668	0.0020	7.80
FSM v. KS1 average	14.38	0.0022	0.0002	5.13

Figure 3.4 Graphical example of interaction pseudo-effect sizes

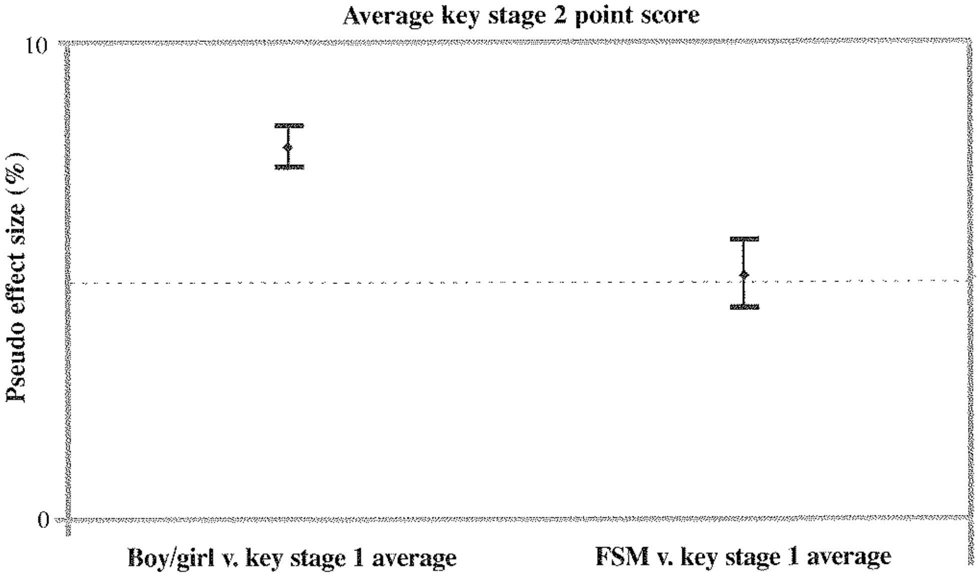
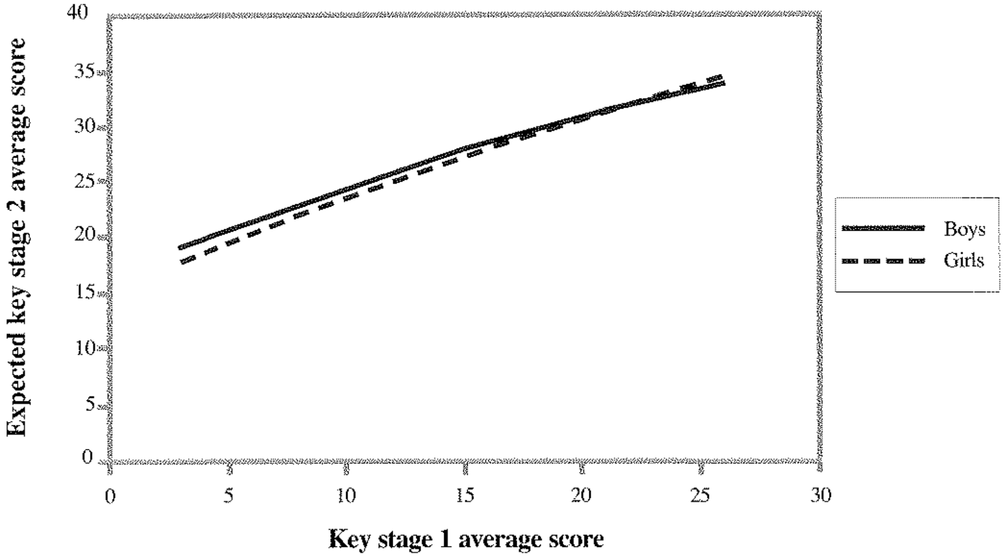


Figure 3.4 shows that both sex and school FSM % have an impact on the link between prior attainment and KS2 outcomes, although the former seems to be the stronger. It might be helpful to illustrate these effects graphically, for example by means of plots such as Figures 3.5 and 3.6.

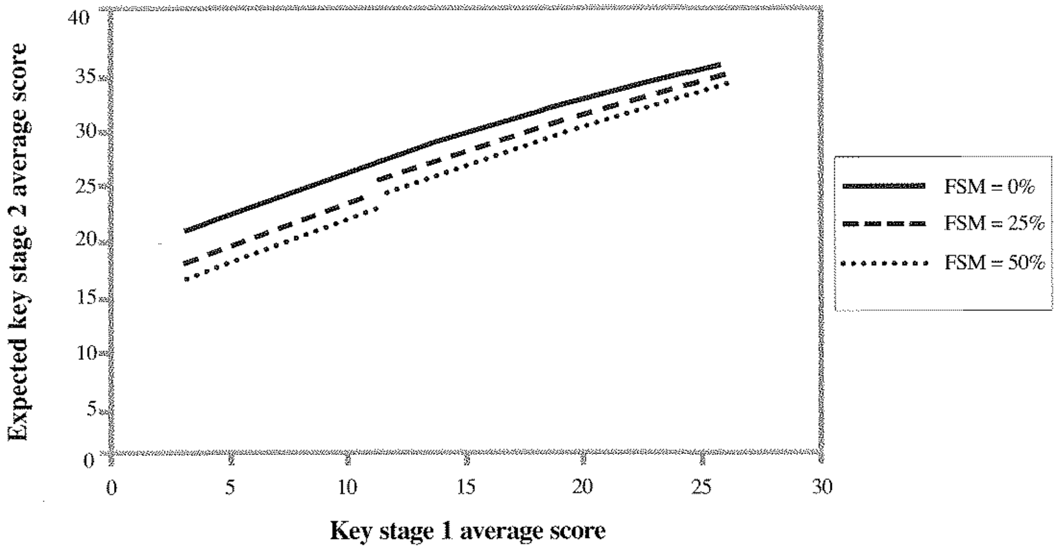
The positive coefficient of SEXINT implies that the relationship with prior attainment is 'steeper' for girls than for boys, as illustrated in Figure 3.5.

Figure 3.5 Average key stage 2 score versus key stage 1 for boys and girls



The positive coefficient of FSMINT implies that the relationship with prior attainment is 'steeper' for higher values of FSM, as illustrated in Figure 3.6.

Figure 3.6 Average key stage 2 score versus key stage 1 by % eligible for FSM



3.2.5 Adjusted coefficients

We have developed a range of indictors derived from the original coefficients, but all expressed in dimensionless terms. These can be quite useful to compare the relative strengths of the different effects and allow us to relate to the concept of effect sizes and the related literature in this area.

But do the 'consumers' of our research really want to think in dimensionless terms all the time? Is there not a case for saying that they are interested in the actual impact which different factors have on pupils' attainment? If this is true, then it makes more sense to express the results in terms of the original units, but in comparable terms which take account of the underlying ranges of the different factors. Let us define an 'adjusted coefficient':

$$\begin{aligned} a &= B && \text{if } X \text{ is binary,} && (5) \\ &= B^*\surd 2s && \text{if } X \text{ is not binary.} \end{aligned}$$

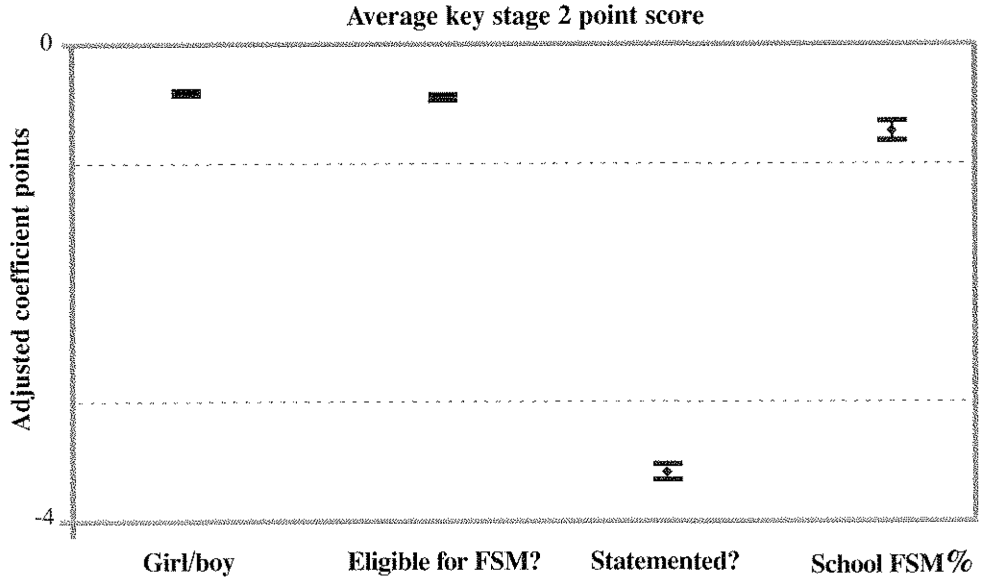
Then a tells us the expected change in Y (in whatever units it is measured in) for an 'average switch' in the values of X. Table 3.4 shows these values for our example case study – note that the units of Y are in key stage 2 point scores (6 per level).

Table 3.4 Adjusted coefficients for selected background variables related to key stage 2 average point score outcome in 2002

Description	Standard deviation	Range	Coefficient	Standard error of coefficient	Adjusted coefficients
Girl/boy	0.5	0,1	-0.411	0.007	-0.411
Eligible for FSM?	0.37	0,1	-0.436	0.010	-0.436
Statemented?	0.15	0,1	-3.593	0.030	-3.593
School FSM %	14.38	0,100	-0.036	0.002	-0.724

Figure 3.7 shows these values; it is directly comparable to Figure 3.3, except that the vertical axis has different units.

Figure 3.7 Graphical example of adjusted coefficients



3.3 The choice of units for presenting results

In the previous section we explored a number of possible dimensionless quantities that might be useful for presenting and comparing the impacts of different background factors on the outcome, and ended by considering the possibility of using a measure with the same units as the outcome, on the basis that this might have a more direct meaning for the consumers of our analysis. However, to a large extent this depends on the exact units in which the outcome is measured – some are likely to be more meaningful than others.

In much of this work, the outcome variable relates to the performance of individuals in some kind of assessment. The scales which can be used to measure this include:

- total test score
- standardised score (adjusted for age or otherwise)
- national curriculum levels
- point scores derived from national curriculum levels (6 points per level)
- GCSE points based on grades (8 for A* down to 1 for G) – either total, average or specific subject based.

Of these, total test score is quite hard for a general audience to relate to, without a great deal more detail about the range and difficulty of the test and it is difficult to make impact measures expressed in these units accessible. Standardised score points (if normalised to a mean of 100 and standard deviation of 15) are easier to understand, as many people can relate them to their idea of 'IQ' and to the concept that 95% of individuals lie in the range 70 to 130. However, without more work it is not immediately apparent what an impact measure of, say, 3 standardised score points might mean in educational terms.

With measures related to the national curriculum we are on stronger ground, as there is a shared understanding within the English education system of what is understood by a 'level'. The point score equivalents (6 to a level) are tied into the same system. In Table 3.4, for example, the girl/boy difference coefficient is -0.411, expressed in the latter units. Expressed in levels, this is -0.069, and this represents the progress that girls are making relative to boys during KS2. It is still not entirely clear, however, how big a difference this really is in educational terms.

One possible approach is to think in terms of nominal months of progress. The original report, on which the national curriculum levels were based (DES, 1987), suggested that the average pupil would progress through a level in two years. This gives us a conceptual yardstick, even if the current amount of progress is undoubtedly different in practice. Let us set out the following equivalences:

1 level = 6 points = 24 'TGAT months'.

So the girl/boy difference of -0.411 becomes -1.64 'TGAT months', and it might be acceptable to interpret this along the lines of: 'During KS2, boys make about one and a half months more progress on average across the core subjects than girls'. Table 3.5 shows the adjusted coefficients for our example expressed on three scales: point scores, levels and months.

So if we were asked about the impact of these factors on pupils' progress through KS2, we might be able to express the results in terms which were readily accessible. For example, statemented pupils make about 14 months less progress than would have been expected; or school-level FSM makes a difference of about three months on progress.

Table 3.5 Adjusted coefficients for selected background variables related to KS2 average point score outcome in 2002, expressed on different scales

	Adjusted coefficients		
Description	**Points**	**Levels**	**Months**
Girl/boy	-0.411	-0.068	-1.64
Eligible for FSM?	-0.436	-0.073	-1.75
Statemented?	-3.593	-0.599	-14.37
School FSM %	-0.724	-0.121	-2.90

If we decide that this metric is the right one for displaying results, can we use it for other outcome scales? It is not clear at the moment how to do this for GCSE outcomes, but there is a possible approach for age-standardised scores. Suppose we have determined an 'impact measure' equivalent to 5 standardised score points – how does this equate to months of progress? Table 3.6 below is an example of (part of) an age-standardisation table, which enables age-standardised scores to be computed for any combination of 'raw score' (down the side) and age in years and completed months (across the top).

Table 3.6 Example of age-standardisation table

	10.05	10.06	10.07	10.08	10.09	10.10	10.11	11.00	11.01	11.02	11.03	11.04	11.05	11.06
7	90	89	89	88	88	87	87	87	86	86	85	85	84	84
8	92	91	91	90	90	89	89	89	88	88	87	87	86	86
9	94	94	93	93	92	92	91	91	90	90	89	89	88	88
10	96	96	95	95	94	94	93	93	92	92	91	91	91	90
11	98	98	97	97	96	96	96	95	95	94	94	93	93	92
12	101	100	~~100~~	~~99~~	~~99~~	~~98~~	~~98~~	~~97~~	~~97~~	~~96~~	~~96~~	95	95	94
13	103	102	102	102	101	101	100	100	99	99	98	98	97	97
14	105	105	~~104~~	~~104~~	~~104~~	~~103~~	~~103~~	~~102~~	~~102~~	~~101~~	~~101~~	100	100	99
15	108	108	107	107	106	106	105	105	104	104	103	103	103	102
16	111	110	110	109	109	109	108	108	107	107	106	106	106	105
17	114	113	113	113	112	112	111	111	111	110	110	109	109	109
18	117	117	117	116	116	116	115	115	115	114	114	114	113	113
19	121	121	121	121	121	121	120	120	120	119	119	119	119	118
20	132	132	132	132	131	131	131	131	131	131	131	131	130	130

To investigate the impact of a 5-point change in standardised scores, we might adopt the following procedure:

1 locate a suitable age-group, for example the lowest appropriate age for the individuals concerned (in our example we have taken 10.06)

2 locate the cell of the table with an age-standardised score of 100 at this age

3 move along this row (keeping the raw score constant) to find a cell with value 95 (in our example 11.04)

4 compute the difference in ages (in our case, 10 months).

We could repeat the procedure, for example starting at 105 and ending at 100. In Table 3.6 this would also give a 10 month gap. Therefore, we might argue that a 5-point difference in standardised scores was approximately equivalent to a 10-month difference in ages.

There are various problems with this approach. One is that age-standardisation tables are not always as linear in form as shown in Table 3.6 and the results obtained may vary according to the start and end points chosen. More fundamentally, however, it makes the assumption that changes in performance across different age groups are equivalent to changes over time due to maturation. This is quite a strong assumption to make, although if we are prepared to accept it this allows us to use the above technique to generate approximate progress measures using months.

The relationship between measures computed in this way from age-standardised scores and those derived from national curriculum scores or levels via the 'TGAT assumption' has not been researched. If it is felt that this is a suitable metric in which to present complex modelling results it may be worthwhile carrying out further investigations.

3.4 Summary and conclusions

In this paper I have explored a number of possible ways of presenting the outcomes of complex modelling processes, based on the fixed part coefficients from a multilevel analysis. These have included:

- **'raw' coefficients**: directly derived from the model results, in outcome units/background unit, and not directly comparable with each other for variables which are in different units
- **normalised coefficients**: dimensionless, scaled by the standard deviation of both outcome and background variable, and directly comparable in terms of the impact of the given factors across the whole population

- **pseudo effect sizes**: dimensionless, scaled by the outcome standard deviation and the 'average switch' between values for the background variable, and directly comparable in terms of the impact of each factor for those whom it affects
- **interaction pseudo effect sizes**: dimensionless, scaled by the main effect coefficient of the more basic factor and the 'average switch' between values for the other factor
- **adjusted coefficients**: equal to the raw coefficient for binary variables, and scaled by the 'average switch' between values for non-binary variables, and in outcome units – comparable across background variables in terms of the impact of each factor for those whom it affects.

Graphical ways of displaying these measures, including their confidence intervals, have been demonstrated. These 'star wars plots' enable the user to see at a glance which factors have the largest positive or negative effect, and whether or not they are statistically significant. Even if this particular layout is not regarded as the best, some form of graphical presentation is worth exploring for any complex model.

A further issue is the question of the units in which results could be reported. These might be dimensionless (i.e. similar to effect sizes) or could be in other units which are more directly interpretable by policy-makers and other consumers of this kind of analysis. A prime candidate seems to be the concept of 'months of progress', either derived from national curriculum levels via the 'TGAT model' of progress, or from age-standardised scores via the age allowance built into the standardisation table. Personal experience has shown that this measure is popular with policy-makers, but it remains to be seen if this is true across the whole range of those interested in the results of our modelling.

This paper has explored some important areas, although it cannot claim to be the last word on any of them. Perhaps the time has come for educational researchers and statisticians to more towards a consensus about how we report the results of complex modelling so that our wider and less technically-focused audience can get to grips with them.

Acknowledgements

Thanks to my colleagues who have commented on this paper, including Paula Hammond, Michela Gnaldi and Tom Benton. Thanks also to Professor Harvey Goldstein for his incisive queries about an earlier draft. If the paper has not significantly improved following their comments this is not their fault but mine.

References

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.

DEPARTMENT OF EDUCATION AND SCIENCE (1987). *Task Group on Assessment and Testing: a Report*. London: DES.

GORARD, S. (2003). 'What is multi-level modelling for?' *British Journal of Educational Studies*, **51**, 1, 46–63.

JESSON, D. (2003). 'Methods for evaluating school performance and their impact on debate about specialist schools', *Research Intelligence*, **82**, 27.

SCHAGEN, I. and BENTON, T. (2003). 'The relationships between school and pupil background factors and progress from key stage 1 to key stage 2.' Paper presented at the British Educational Research Association Annual Conference, Heriot-Watt University, Edinburgh, 13 September [online]. Available:http://www.nfer.ac.uk/research/papers/ISTB bera03.doc [4 March, 2004].

SCHAGEN, I. and HUTCHISON, D. (2003). 'Adding value in educational research - the marriage of data and analytical power', *British Educational Research Journal*, **29**, 5, 749–65.

4 The use of effect sizes: two examples from recent educational research

Steve Strand

4.1 Introduction

What is 'effect size'? At its simplest, the effect size (d) is just a standardised measure of the difference in the mean scores of two groups. It is calculated as the difference between the two means divided by the pooled standard deviation. (See Coe, 2002 for a discussion on issues around the calculation of the 'pooled' standard deviation). Its principal application to date has been in meta-analysis, which seeks to combine and compare estimates from different studies. For example we may wish to compare gender differences on the 60–140 scale of a standardised test score, with differences on the 9 point (A*–U) scale of a GCSE examination grade, with differences on a 0%–100% measure of attendance at school, and with differences on a 48–240 scale on a 48 item Likert type questionnaire assessing attitudes to learning. The effect size provides us the standardised index to make meaningful comparisons across these different measures.

Effect sizes are not a new concept, although they only became prominent with the rise of meta-analysis in the early 1970s. There has been a move to make their use more widespread and the American Psychological Association (APA) has advocated their use since 1994. For example, the following is taken from the manuscript submission guidelines of the *Journal of Educational Psychology.*

> *Appropriate indexes of effect size or strength of relationship should be incorporated in the results section of the manuscript (see pp. 5, 25–26 of the APA Publication Manual). Information that allows the reader to assess not only the (statistical) significance, but also the magnitude of the observed effects or relationships, clarifies the importance of the findings.*
>
> (www.apa.org/journals/edu/submission.html)

Despite the above, effect sizes do not appear to be routinely quoted within the educational and psychological literature. A recent paper by Coe (2002) has reviewed issues around the calculation, use and interpretation of effect sizes in educational and social science research and argued for their wider use.

The specific aim of this paper is to illustrate, with examples from two recent papers, the limitations of interpreting results solely on the basis of traditional statistical significance testing, and to show how effect sizes can improve interpretation. The two examples show traditional significance testing is limited or insufficient for purpose where:

- sample sizes are extremely large, and consequently even 'small' differences may be statistically significant
- we want to estimate the relative magnitude of the effects of a range of independent variables (e.g. in a multiple regression equation) that all exceed conventional levels of statistical significance, for example all are p<0.01.

A further aim of the paper is to broaden the discussion of 'effect size' beyond measures of central tendency. The concept of the variance ratio will be illustrated to show how to assess the magnitude of group differences in score variance, in the same way that 'd' assesses the magnitude of group differences in mean scores.

4.2 Example 1: Large sample sizes – sex differences in cognitive abilities test scores (Strand, 2003)

4.2.1 The dataset

There continues to be debate on the extent, or even existence, of sex differences in the mean level and variability of cognitive ability test scores (Lynn, 1994, 1998; Mackintosh, 1996). However the debate suffers from a lack of studies that:

- are based on large nationally representative population samples
- distinguish between educational attainment and reasoning abilities
- disaggregate separate reasoning abilities, as distinct from IQ
- are drawn from outside the US
- analyse recent test data (i.e. administrations within the last 30 years).

In contrast to the above, Strand (2003) reports the Cognitive Abilities Test (CAT) scores of a nationally representative UK sample of over 320,000 pupils aged 11–12 years assessed between September 2001 and August 2003 on the recently UK standardised CAT3, which includes tests of Verbal Reasoning (VR), Quantitative Reasoning (QR) and Non-

Verbal Reasoning (NVR). The substantive research question is: what is the extent (if any) of sex differences in cognitive abilities test scores?

4.2.2 Results

Table 4.1 presents for boys and girls separately the sample size, mean and standard deviation for standard age scores on each of the three CAT batteries and for mean CAT score (the mean of the three separate batteries).

Table 4.1 Mean, standard deviation and sample size for boys and girls on CAT3 Level D (from Strand, 2003)

CAT score	Statistic	Boys	Girls	Significance of difference	Effect Size (Variance Ratio)
Verbal	Mean	98.4	100.6	P<.0001	0.15
	SD	15.1	14.5	P<.0001	(1.09)
	N	158,093	158,457		
Quantitative	Mean	99.4	98.9	P<.0001	-0.03
	SD	15.0	13.8	P<.0001	(1.18)
	N	157,862	158,406		
Non-Verbal	Mean	99.7	100.2	P<.0001	0.03
	SD	14.8	13.9	P<.0001	(1.13)
	N	157,830	158,299		
Mean CAT score	Mean	99.1	99.9	P<.0001	0.05
	SD	13.5	12.7	P<.0001	(1.13)
	N	156,556	157,258		

Note: Positive effect size indicates female mean greater than male mean. VR greater than 1 indicates male variance greater than female variance.

4.2.3 Sex differences in mean scores

I believe the typical first impression on looking at the data in Table 4.1 are:

- all sex differences are **highly statistically significant**, p<0.0001.

When scanning a paper this may be all we take in, although the more diligent reader may shortly thereafter note that:

- three of the four mean differences are **less than one score point on a scale with a SD of 15** (the sex differences are only 0.5, 0.6 and 0.8 standard score points on NVR, QR and mean CAT respectively).

For many readers I suspect there may be no 'dissonance' between these two observations. Statistics training for psychology undergraduates can sometimes focus almost exclusively on the statistical significance of the outcome (the *p* value) as a way of negotiating the complexities of sample sizes, variances, t and f ratios etc. However given the extremely large sample sizes in Table 4.1 almost any sex difference will be statistically significant; *p* values are therefore a poor guide to the wider psychological or educational significance of the results.

Because the standard age score (SAS) metric (mean=100, SD=15) is widely used in educational tests, it may be that informed readers will not over-interpret the statistical significance of the results. However, it is more common that results are reported on an unfamiliar or non-standard metric (e.g. test raw scores, % correct, factor loadings etc). Further, even though the SAS scale may be familiar, how should we interpret the 'moderate' sex difference of 2.2 SAS points on the Verbal Reasoning battery?

In short, whatever the level of statistical significance, we need to explicitly consider the effect size (d), defined as the difference between the mean scores for boys and girls divided by the pooled standard deviation, to estimate the magnitude of the results. Typically, following Cohen (1977) effect sizes smaller than 0.20 are considered very small and treated as insignificant; effect sizes in the range 0.20–0.50 are considered 'small' but worth noting; 0.50–0.80 are considered medium and above 0.80 is considered large.

The effect size for verbal reasoning is therefore very small (d = 0.15) and the sex difference in Quantitative, Non-verbal and Mean CAT score are clearly negligible (d = +/–0.05).

4.2.4 Sex differences in score variance

Effect size research has concentrated on measures of central tendency, such as the mean, paying little attention to dispersion as an outcome in its own right. For example, because the developers of modern meta-analytic techniques were concerned exclusively with cumulating results of comparisons between groups in mean scores, this became the sole concern in all subsequent meta-analytic research (Feingold, 1992). However we can also ask how groups differ in the SD or variance of their scores. Is the variance of scores greater for one group than another?

The variance ratio is formed by dividing the variance (SD squared) for one group by the variance for the second group. In the case of sex differences, male variance is traditionally divided by female variance so a ratio>1 indicates the variance is greater for boys than for girls, while a ratio<1 indicates greater variability in the scores of girls. When viewed as a descriptive statistic, Feingold (1992) suggests a variance ratio above 1.10 (or below 0.9) is probably the smallest meaningful effect.

In these terms, sex differences in variability are very close to the threshold for verbal reasoning and exceed the threshold for both non-verbal and quantitative reasoning (see Table 4.1). In percentages terms boys' scores are 9% more variable than girls on verbal reasoning, 13% more variable on non-verbal reasoning and 18% more variable on quantitative reasoning. Figure 4.1 presents a graphical illustration of the percentage of boys and girls within each of nine (stanine) standard age score bands.

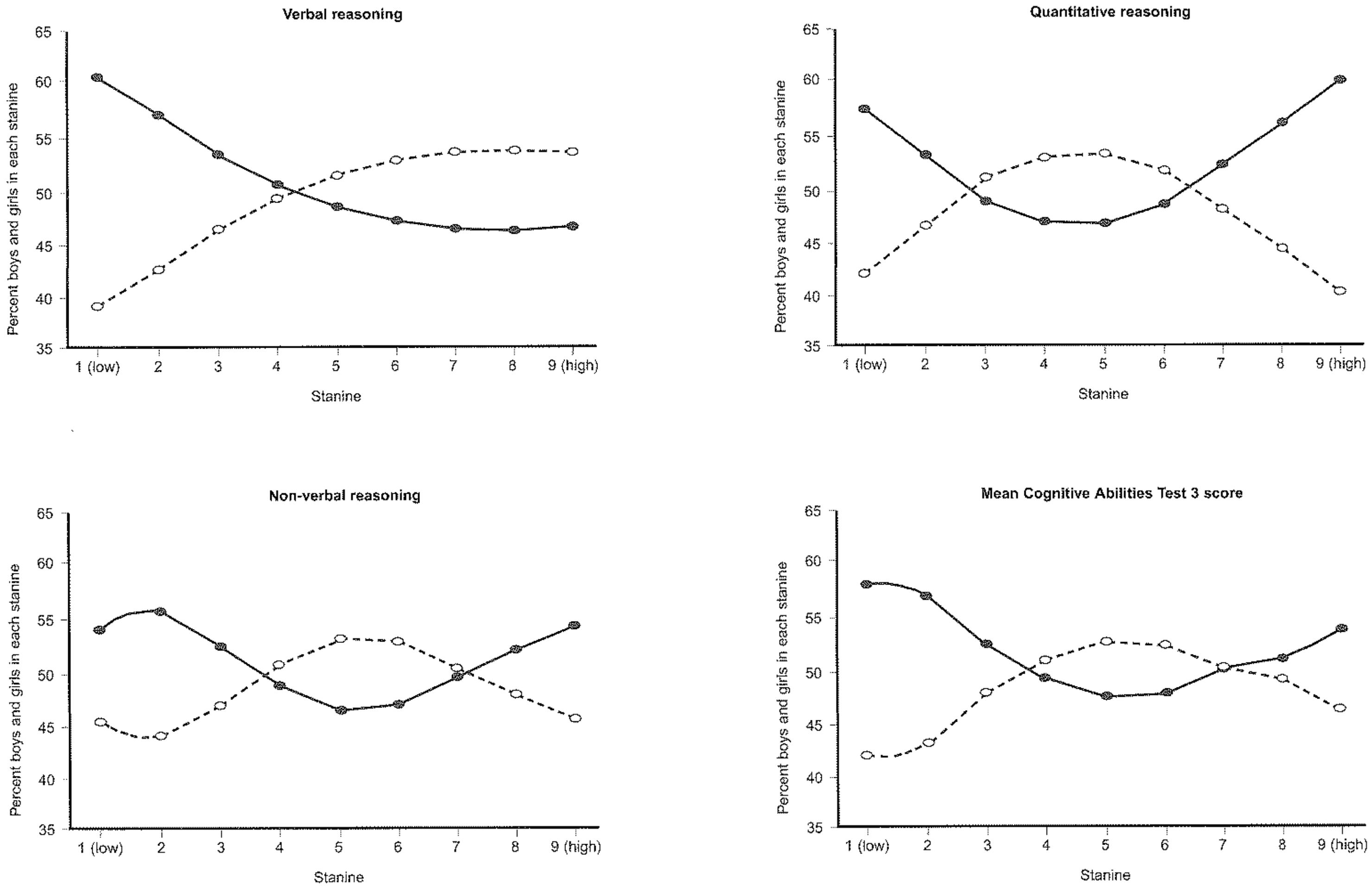
The differences in variability between the sexes are not huge. Sixty per cent of the pupils scoring in the bottom 5% of the VR range, and in the top 5% of the QR range, were boys, giving a ratio of 1.5:1. This indicates that three of every five pupils identified with these extreme scores will be boys. Differences in the top and bottom 5% of scores for NVR are slightly less extreme, with around 55% boys, or a ratio of 1.25:1, indicating that five of every nine pupils identified at these extremes are likely to be boys.

4.2.5 Example 1 conclusions

The conclusions with respect to methodology from this example are that:

- statistical significance testing is important, but not sufficient to assess the psychological/educational significance of results with very large samples
- measures of effect size (d) offer a direct indicator of the magnitude of the group 'effect', independent of sample size, and are therefore more useful
- measures of group differences in score variance (VR) are just as important as differences in mean score (d).

Figure 4.1 Percentage of boys and girls within each stanine score band



Verbal reasoning

Quantitative reasoning

Non-verbal reasoning

Mean Cognitive Abilities Test 3 score

The substantive conclusions of the research are that:

- the lack of substantial sex differences in reasoning test scores suggests that explanations of sex differences in public examinations results must look beyond conceptions of 'ability'
- despite the prominent media and government focus on the 'gender gap', educators must be careful to avoid general conceptions of boys as underachievers. The current study suggests that boys are slightly more likely to be over-represented relative to girls at the high as well as the low extremes of reasoning ability
- the greater variability in boys' reasoning scores may explain to some extent their greater representation within populations with Special Educational Needs and among those who fail to achieve any GCSE or equivalent passes. However, boys do not appear to be over-represented at the higher (A*) end of GCSE performance. To this extent, it may be valid to speak of a degree of underachievement at GCCE, particularly among more cognitively able boys.

4.3 Example 2: Establishing the relative impact of a range of equally statistically significant factors – does pupil mobility (changing school) affect educational progress between age 4 and age 7?

4.3.1 Background

As an example we take a study by Strand (2002). The study was concerned to determine the impact of pupil mobility on pupils' educational progress between age four and age seven, while controlling for a wide range of other pupil level and school level variables. The data was drawn from 6,300 pupils attending 56 primary schools in an inner London LEA. Pupils' completed baseline assessment on entry to reception class at age four and were tracked until they completed national end of key stage 1 (KS1) tests at age seven some three years later. A wide range of background data was collected on the pupils, such as their age, sex, entitlement to Free School Meals (FSM), ethnic group, stage of fluency in English and stage of Special Educational Need (SEN). The paper sought to determine if there were any differences in educational progress between 'stable' pupils, defined as those who had attended the same school for the whole of the relevant phase, and 'mobile' pupils, defined as those who joined the school part way through the phase, while simultaneously controlling for all relevant pupil background characteristics.

4.3.2 Results

Table 4.2 presents a summary of the main results. The first four columns show the regression coefficients. The table has been simplified for clarity by showing only statistically significant coefficients, where the regression coefficient is more than twice its standard error ($p<0.05$). We will consider as an example the results for the end of KS1 mathematics test. There is a significant impact on age seven mathematics score of:

- age four baseline test score
- mobility (pupils who had been in the same school for the whole three years made more progress than pupils that joined their school part way through the phase)
- sex (girls made less progress than boys)
- socio-economic disadvantage (pupils entitled to a FSM made less progress than those not entitled)
- SEN stage (the higher the stage the slower the progress, with the greatest negative impact for stages 3–5, then for stage 2 and then for stage 1)
- EAL (greater progress made by EAL pupils who were fully fluent in English (stage 4) than monolingual pupils)
- ethnic group (African & Caribbean pupils made less progress, and Chinese pupils made more progress, than English, Scottish, Welsh & Northern Irish (ESWNI) pupils)
- interactions between ethnic group and FSM, and between sex and ethnic group (e.g. Pakistani girls made substantially less progress than Pakistani boys)
- school level aggregates, such as the % of the cohort entitled to FSM (less progress made in schools with a high proportion of pupils with socio-economic disadvantage).

Table 4.2 Multi-level regression analysis of progress during key stage 1 (taken from Strand, 2002)

Fixed Effect	Regression Coefficient				Effect Size			
	reading	writing	average maths	KS1 score	reading	writing	average maths	KS1 score
Constant	2.864	2.463	2.786	2.704	–	–	–	–
Baseline test score	0.277	0.227	0.313	0.273	0.76	0.70	0.92	0.86
Mobile	–	–	-0.085	-0.055	–	–	0.12	0.09
SEN stage 1	-0.528	-0.387	-0.334	-0.417	0.72	0.60	0.49	0.66
SEN stage 2	-0.737	-0.514	-0.501	-0.586	1.01	0.80	0.73	0.92
SEN stages 3-5	-0.791	-0.680	-0.595	-0.690	1.08	1.05	0.87	1.09
Sex	0.037	0.047	-0.164	-0.027	0.05	0.07	0.24	0.04
Free School Meal (FSM)	-0.157	-0.140	-0.100	-0.132	0.22	0.22	0.15	0.21
EAL: complete beginner	-0.538	–	–	-0.255	0.74	–	–	0.40
EAL: considerable support	-0.176	–	–	-0.102	0.24	–	–	0.16
EAL: some support	–	–	–	–	–	–	–	–
EAL: fully fluent	0.111	0.089	0.126	0.109	0.15	0.14	0.18	0.17
African	–	–	-0.164	–	–	–	0.24	–
Caribbean	-0.079	–	-0.169	-0.098	0.11	–	0.25	0.15
Chinese	0.197	0.254	0.201	0.220	0.27	0.39	0.29	0.35
FSM * African	–	0.132	0.092	0.107	–	0.20	0.13	0.17
FSM * Caribbean	0.166	0.084	–	0.088	0.23	0.13	–	0.14
FSM * any other group	0.150	–	–	–	0.21	–	–	–
FSM * Indian	0.205	–	–	0.149	0.28	–	–	0.23
Sex * Pakistani	–	–	-0.172	–	–	–	0.25	–
School mean %FSM	-0.003	-0.002	-0.003	-0.003	0.16	0.12	0.18	0.19
SD of independent variable	0.730	0.646	0.684	0.635	–	–	–	–
% variance at school level	7.0%	6.8%	9.6%	11.1%	–	–	–	–

Notes: Only significant coefficients (at least p<0.05) are shown. The interaction between mobility and FSM was tested and was not significant for any subject. SD = standard deviation.

The multiple regression establishes that all these variables have an *independent* and *statistically significant* impact on progress, but what is their relative magnitude? The regression coefficients are not themselves directly comparable. There are four types of significant variables:

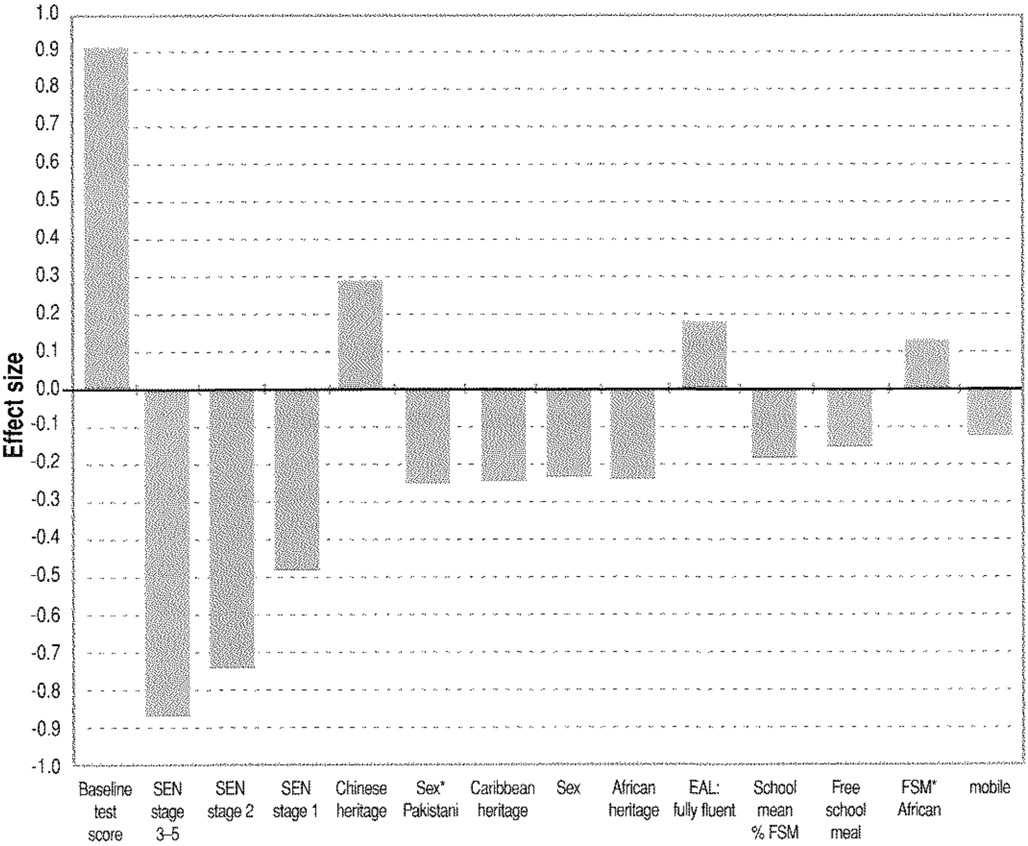
1. dichotomous variables e.g. sex, entitlement to FSM, mobility. We have also translated some nominal measures (such as ethnicity) and ordinal measures (such as stage of English fluency 1–4 and stage of SEN 1–5) into a series of dichotomous dummy variables
2. continuous variables (e.g. age four baseline score), which may or may not have been 'centred' on the population mean
3. interaction terms, e.g., the interaction between sex and Pakistani heritage, or between FSM entitlement and African heritage
4. school level aggregate variables (e.g. percentage entitled to free school meals, mean prior attainment score etc.), which again may or may not have been 'centred' on the grand mean.

Effect sizes offer a means of placing these regression coefficients on a standard index. Following Tymms *et al.* (1997) the effect size of a dichotomous variable is calculated as the regression coefficient divided by the outcome SD. For a continuous variable, the effect size is calculated as the regression coefficient multiplied by 2* the variables' SD, divided by the outcome SD, so that the effect size corresponds to the difference between predicted scores one SD above and one SD below the mean.

- Thus for a simple dichotomous variable such as mobility, the effect size is simply –0.085 / 0.684 = –0.12.
- For a continuous variable such as baseline test score, which was centred around the grand mean, the standard deviation has already been standardised to 1 and the calculation therefore reduces to (0.313 * 2) / 0.684 = 0.92.
- Interactions between dummy variables are also assessed as dichotomous variables. Thus the effect size for Sex * Pakistani is –0.172 * / 0.684= –0.25.
- For the school aggregate measures, the general principle for continuous variables was applied. The standard deviation at the school level of the variable '% entitled to free school meals' was 20%, thus the effect size was (–0.003 * 2 * 20) / 0.684 = –0.18.

This process resulted in the effect sizes included in Table 4.2 and shown graphically in Figure 4.2.

Figure 4.2 Effect size for pupil and school level measures on key stage 1 mathematics test outcomes



4.3.3 Example 2 conclusions

Using effect sizes allows us to compare the different variables on a standard index. For example the regression coefficient for the school composition variable %FSM is very small (–0.003) compared to the coefficient for EAL fully fluent (0.126). However, they both have the same impact on pupil progress, with an absolute value for each effect size of 0.18.

In relation to pupil mobility, we can conclude that while mobility has a statistically significant impact on pupil progress in mathematics, the magnitude of the effect is minimal, compared to the impact of other measured pupil background and school context factors.

4.4 Overall conclusions

The above examples show how effect sizes have aided the interpretation of the results from these studies. Hopefully, these examples might encourage researchers to use effect size more frequently in evaluating their findings. However the following cautions are urged.

First, effect sizes can exceed d>0.20, even when the effect is not statistically significant. It is therefore suggested that statistical significance should be used as a filter and only statistically significant variables should be evaluated for their effect size.

Second, 'd' and 'VR' statistics can be directly compared to previous studies quoting d and VR statistics, but care is needed to ensure the comparability of the measures in such studies. For example, in discussing sex differences in verbal ability, Hyde and Linn (1988) observe that

> *'verbal ability' has been used as a category to include everything from quality of speech in two year olds, to performance on the Peabody Picture Vocabulary Test at age five years, to essay writing by high school students, to solutions to anagrams and analogies.*
>
> (Hyde and Linn, 1988)

It is therefore important to consider carefully the actual measures employed in other studies.

Finally, there has been considerable debate over the interpretation of effect sizes. Cohen (1997) considers d values of 0.2 – 0.49 'small', d of 0.5 – 0.79 as 'medium' and d of 0.8 or above as 'large'. Rosenthal and Rubin (1982) on the other hand have introduced the Binomial Effect Size Display (BESD) as a means of determining the practical significance of an effect size and they have argued that many effect sizes that seem to be small are actually large in terms of practical significance. For example, an effect size of 0.4, and therefore small in relation to Cohen's criteria, would translate into a decrease in the death rate from 60 per cent to 40 per cent when measuring the success of a treatment for cancer, something that undoubtedly has practical significance. It is therefore important to adopt a relative rather than absolute perspective, judging the effects in relation to the particular outcome being studied and compared to other interventions.

References

COE, R. (2002). 'It's the effect size, stupid: what effect size is and why it is important.' Paper presented at the British Educational Research Association Annual Conference, Exeter, 12–14 September [online]. Available: http://www.leeds.ac.uk/educol/documents/00002182.htm [4 March, 2004].

COHEN, J. (1977). *Statistical Power Analysis for the Behavioural Sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates.

FEINGOLD, A. (1992). 'Sex differences in variability in intellectual abilities: a new look at an old controversy', *Review of Educational Research*, **62**, 61–84.

HYDE, J.S., and LINN, M.C. (1988). 'Gender differences in verbal ability: a meta-analysis', *Psychological Bulletin*, **104**, 1, 53–69.

LYNN, R. (1994). 'Sex differences in intelligence and brain size: a paradox resolved', *Personality and Individual Differences*, **17**, 2, 257–71.

LYNN, R. (1998). 'Sex differences in intelligence: a rejoinder to Mackintosh', *Journal of Biosocial Science*, **30**, 529–32.

MACKINTOSH, N.J. (1996). 'Sex differences and IQ', *Journal of Biosocial Science*, **28**, 559–71.

ROSENTHAL, R., and RUBIN, D. B. (1982). 'A simple, general purpose display of magnitude of experimental effect', *Journal of Educational Psychology*, **74**, 2, 166–9.

STRAND, S. (2002). 'Pupil mobility, attainment and progress during key stage 1: a case study in cautious interpretation', *British Educational Research Journal*, **28**, 1, 63–78.

STRAND, S. (2003). 'Sex differences in cognitive abilities test scores: a national picture.' Paper presented at the British Educational Research Association Annual Conference, Heriot-Watt University, Edinburgh, 11–13 September.

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.

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:

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

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

These may be combined to give a single equation:

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

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:

$$\Delta = \frac{\overline{X}_{Exp} - \overline{X}_{Cont}}{SD_{pooled}} \tag{4}$$

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:

$$y_{ij} = \beta_0 + \beta_1 + u_j + e_i \tag{5}$$

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:

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

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

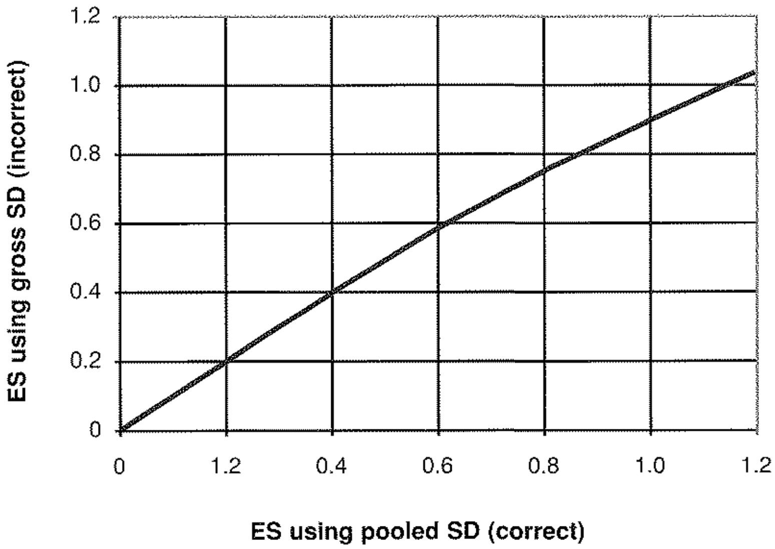
Variables often appear in multilevel models simply as controls. That is to say, they are there to improve the model or because there is an inherent interest in them and not because they measure school differences *per se*.

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



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:

$$\Delta = 2zr_{xy}(1 - r^2_{xy})^{-\frac{1}{2}} \tag{8}$$

where:

r is the correlation between variables x and y

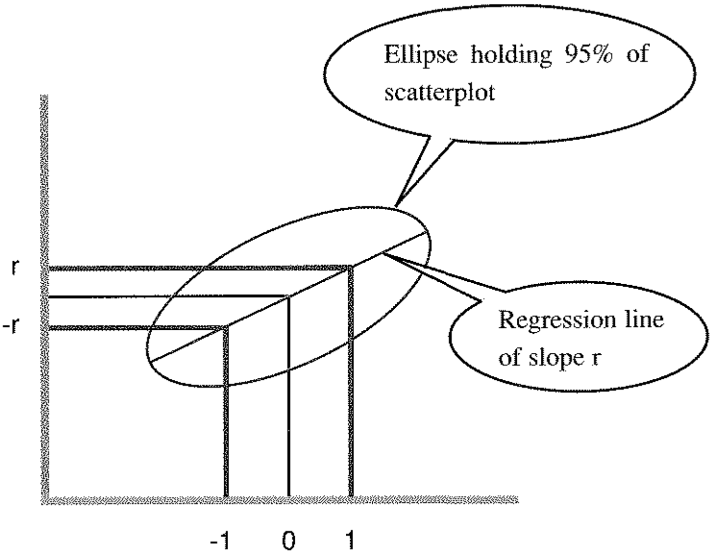
z is the 'unit normal deviate at the pth percentile'

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:

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

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


Ellipse holding 95% of scatterplot
Regression line of slope r
r
-r
-1
0
1

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:

$$\Delta = \frac{2\beta_1}{\sigma_e} \qquad (10)$$

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 β_1* $SD_{predictor}$. The formula is:

$$\Delta = \frac{2\beta_1{}^* SD_{predictor}}{\sigma_e} \qquad (11)$$

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 / √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:

$$\Delta = \frac{2\sigma_u}{\sigma_e} \tag{12}$$

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*√0.029 / √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:

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

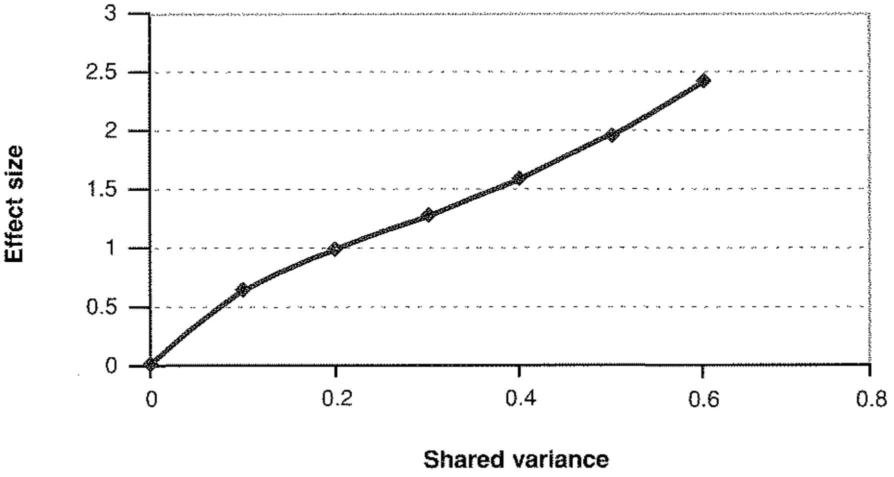
where:

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:

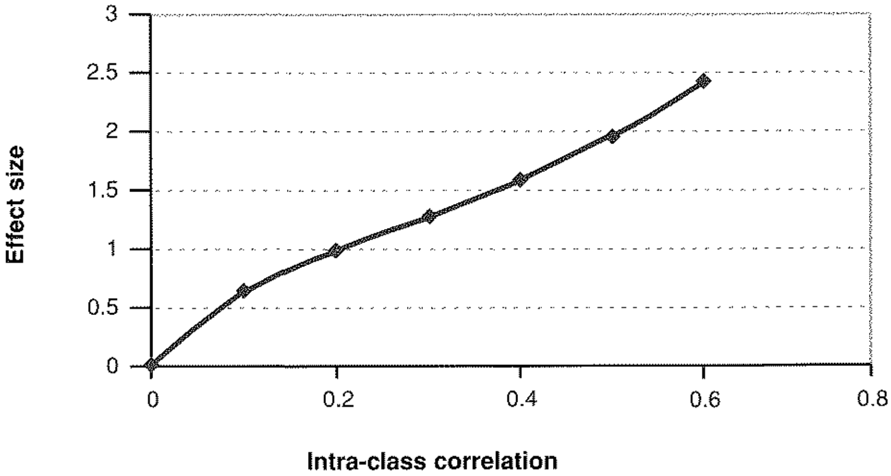
$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2} \tag{14}$$

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.

6 Some observations on the definition and estimation of effect sizes

Harvey Goldstein

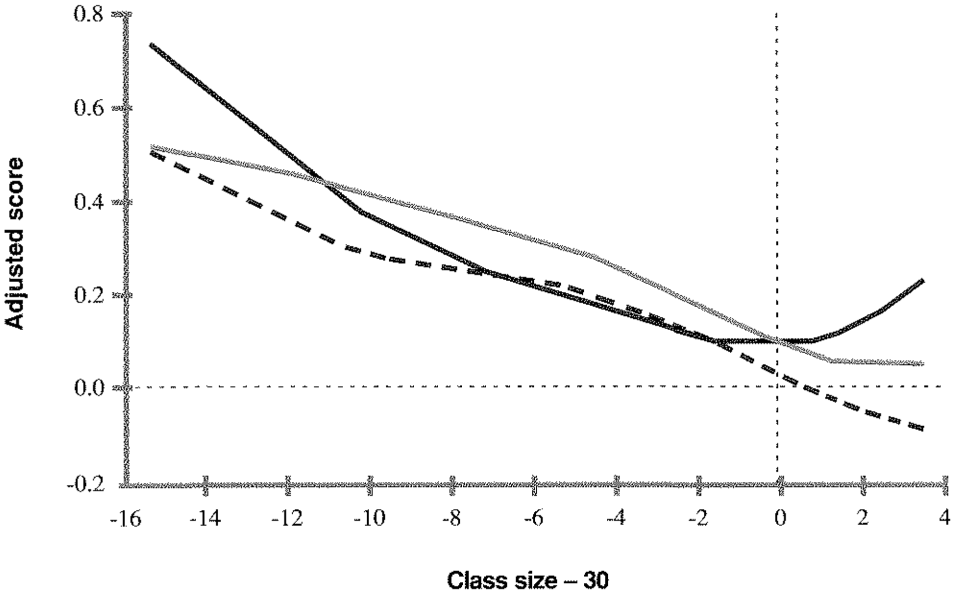
6.1 General considerations

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.

7 Notes as discussant

Trevor Knight

7.1 Background

The 1983 Statistical Bulletin on School Standards and Spending and its 1984 follow-up (13/84) were the first time the then Department of Education and Science (DES) had published for LEAs quantitative measures and consequent conclusions on the relationships between performance and background factors. The analyses contained in the Bulletins were simple stepwise OLS regressions on LEA-level data and represented the reasonable maximum that could be done with national data at the time.

But it was recognised even then that the 'ecological fallacy' was present, and that as a consequence the amount of variation explained at the LEA-level was likely to be hiding significant facts about the variation in pupil progress within and between schools. This feature led to Murray Aitkin, Nick Longford and Harvey Goldstein driving forward the theory of multilevel modelling using ILEA individual pupil data. This led to seminal papers on multilevel modelling – applied locally worldwide. There began to be growing understanding of the practical importance of the multilevel approach for developing interpretations and educational constructs. Academic communities were increasingly using such powerful methods but no national centrally-available datasets were yet available on which to use them.

The Task Group on Assessment and Testing (TGAT) and the introduction of the National Curriculum imposed a system of testing, initially for key stage 1, which by 1996 had extended robust national data to key stage 2 and key stage 3 (though the tests themselves have been subject to regular change). Individual pupil data for key stage 2 and key stage 3 was collected by SCAA (School Curriculum and Assessment Authority, now Qualifications and Curriculum Authority) from the outset and shared with the Department. The Department had responsibility for collecting information on key stage 1 from LEAs.

From 1998, once testing had settled down and become accepted, the construction of data collection mechanisms which facilitated matching pupil test results across time began. SCAA commissioned Carol Fitz-Gibbon from Durham University to consider how 'value-added' measures (based on matched pupil test records, controlling for prior attainment) should develop (Fitzgibbon, 1997). This report was an important landmark in this area. In the interim, following the first schools' White Paper in 1997 of the Labour government, the use of school-level 'benchmarks' based on free school meal (FSM) eligibility rates collected by the Department's Annual Schools' Census (ASC) was introduced, and these continue to be published by DfES/QCA/Ofsted in the Autumn Package and used by Ofsted in Performance and Assessment reports (PANDAs).

Performance Tables were instituted in the early 1990s, based on raw results (colloquially known outside the DfES as 'league tables'). The development of national matched pupil test records allowed Performance Tables to address the value-added issue, and consequently a period of pilots and testing led to national roll-outs of compulsory age value-added (VA) measures, including a key stage 2-GCSE/GNVQ pilot in 2003.

The DfES, after extensive consultation and discussion, constructed school VA measures for Performance Tables on the basis of comparing individual pupil outcomes against that of the 'median pupil' having broadly the same aggregate level of prior attainment. For a variety of practical and policy considerations, a method of VA construction was adopted that was believed to be both open to checking by schools and to be more readily understandable by parents. Construction of measures using a multi-level modelling approach was not then recommended as a feasible option for Performance Tables.

No background information on individual pupils (other than gender) that could potentially have played a part in the VA measures for Performance Tables was available to the Department at that stage. But planning had begun in the late 1990s on the development of a 'Common Basic Data Set' (CBDS) that would govern the information that the DfES and its education partners believed every school and LEA should have commonly available as a minimum to manage their responsibilities.

The first part of the CBDS, the Pupil Level Annual Schools' Census (PLASC), was piloted in 2001 and introduced in full the following year. This contains a number of variables on pupil background – including whether pupils are 'known to be eligible for a free school meal (FSM)', their minority ethnic 'group' and whether they have a statement of educational need (SEN). Pupil home address postcode is also collected, which would allow in time the use on GIS-based area information to indicate pupils' socio-economic circumstances from other sources.

National pupil-matched performance data is linked to individual pupil PLASC returns through the Unique Pupil Number (UPN) as part of DfES' management information system development. This is part of the *Key to Success* initiative with LEAs to improve data collection and transfer arrangements, data interpretation and checking. Pupils' PLASC data linked to their test and exam performance is the basis of the Department's *National Pupil Database* (NPD) which will be a major repository, maintained under comprehensive confidentiality conditions, for the development of pupil, school and LEA performance measures.

One of the first fruits of PLASC/NPD was the launch of an improved 2003 Autumn Package in the form of the Pupil Achievement Tracker (PAT). This was sent to all maintained schools allowing them to input details of their own pupils and their prior attainments and exemplify, using national matched pupil performances expressed in the shape of 'transition matrices', possible levels of performance outcomes. All schools had access to the same information to help them make their contributions to the national literacy and numeracy strategies and targets.

For transparency and ease of interpretation, the transition matrices are based only on prior attainments. They do not contain contextual data, and are not constructed around any educational or statistical model. Furthermore, no information on any uncertainty intervals that will typically apply to estimates of future pupil and school performance is provided.

But by 2002 there was general acknowledgement that the improvements now in place on pupil data availability, allied to the generally accepted view that multilevel modelling offered a powerful technical resource, implied that more refined bases of VA measures could be introduced. A lot of this work has been pioneered by the academic community – The University of London Institute of Education (ULIE) being prominent – but others in higher education working with LEAs are also strongly involved. The Fischer Family Trust (FFT) for example has been providing LEAs with a 'contextualised VA' service, though not yet using multi-level systems, and the results are currently used by DfES in its 'Underperforming Schools' project.

Many have noted that there are at least two sets of school VA measures in the public domain which use different analytical constructs and alternative measures. It has been generally accepted that alternative VA formulations can be used in a complementary way but that the benefits each would bring to discussion, evaluation and policy had to be weighed in the balance.

There is now widespread feeling amongst schools and LEAs (and the academic community) that some contextual measures have a statistically significant and educationally justified impact on the explanation of variation in outcomes at all key stages. Schools and LEAs (and the DfES) are being judged on pupil performance measures, especially the Department's Public Service Agreement (PSA) targets, and that the most appropriate and relevant methods and measures must be seen to being used. The Royal Statistical Society (RSS) Working Party on Performance Indicators has recently issued cogent advice on how national data should be collected, described and used by the Government to advance the public's understanding of information and data analysis.

With this in mind, the DfES set up the Value Added Methodology Advisory Group (VAMAG) as one means of gleaning the advice of key analytical practitioners to help its VA agenda. The group is being informed by the 'intelligent accountability' framework, consonant with the revised OFSTED inspection system. Ministers had expressed their wish to the see improved performance measures widely used.

7.2 Data analysis and presentation issues

The DfES publishes a wide range of statistical information, some of which – Performance Tables and the Autumn Package – explicitly relate to the levels of pupil progress and to the variation in pupil performance conditional on prior attainment. Statistical Bulletins are published which explore, compare and contrast the progress made by pupils controlling for their and their schools' characteristics using information from PLASC. These analyses are population based and, generally, do not contain impact statements or uncertainty profiles.

However, Performance Tables have a section which explains the concept of 'uncertainty' in comparing school value added measures and a technical annex gives guidance to readers on how to interpret those measures nationally and locally. The Tables do not indicate 'confidence intervals' for individual schools: readers do that for themselves in the light of the guidance provided.

Ministers are clear that the Department should use the best available analytical methods and maximise use of available data. There is an acceptance that the sophistication of performance judgements possible has now increased (for the reasons given above) and that careful use of more skilled analysis will enhance evidence-informed policy-making.

However, the advance of the analytical tool-set to improve the range and robustness of performance judgements requires a considerable increase in the understanding and awareness of how these judgements have been created, how they should be interpreted, and how they can be utilised for policy creation and maintenance. There are competing siren calls on methods and methodology each of which will yield different answers in a high stakes environment.

The Department accepts that it now needs to do far more to present comprehensive analyses that describe complex and complicated performance and context relationships in ways that command respect, understanding and acceptance. This requires a large training and development programme from those not conversant with complex data methods throughout and beyond the education sphere.

The concept of effect or impact assessments, and their presentation, is pivotal in developing a clearer appreciation of which factors appear important in developing and evaluating policy, and to which pupil and other groups those effects are or are not important. The presentation of this information, whether by chart, graph or table, is a crucial ingredient in building public understanding of what the Government is doing, how it proposes to do it, and the costs (financial and other) of implementation.

The use of more sophisticated and appropriate methods and the development of describing where significant associations do, or do not, exist, must then be followed by a large suite of published assessments for academic and public review. Presenting significant differences, their relative importance, and which groups are affected, then allows 'data mining' to explore relationships. This can lead to the improvement of educational theory, to hypothecation of potential change to performance given changed conditions, and hence to sets of policy initiatives which can be scaled and ordered in a quantitative way.

Some evaluations commissioned by the Department are already using sophisticated techniques and effect assessments. These will be widened as data and circumstance permits. But it is essential that the practical importance and understanding of the relative importance of factors significant to performance (of pupil and school) is also widened and becomes inculcated within the Department's work.

VAMAG's work is being advanced with the 'effects' issues on-board. There will be publication by the Department of more analyses which put pupil and school performance and variation robustly in the wider context. The Department has made commitment to the use of impact assessments in what it publishes, and in what it uses to inform policy development.

Consequently, it is vital that the educational community makes available the appropriate analytical theory supported by practical examples in ways which are accessible to the Department. These should be supplemented by ways in which the data could be examined further to highlight potential areas for more analysis relevant to current or potential policy concerns.

7.3 Conclusions

- The extent of disaggregated education data has expanded considerably, and is continuing to expand, as datasets are linked through postcodes and other keys.
- There is now a widely available and accepted bedrock of appropriate statistical theory to employ on the widening volumes of disaggregated data, informing the web of educational relationships.
- There are countless analytical examples of sound analytical theory and practice on pupils and school performance matched with high quality presentation on which the Department can draw.
- As it develops its 'intelligent accountability' concepts to assist higher quality policy making, the Department will help schools and LEAs with better information for reflection, improvement and target setting.
- This assistance includes the use and dispersion of more sophisticated analytical methods and creates within the Department a challenge to policy making and to the central Government's relationships with the education world.

- The Department is committed to working with partners to use more robust methods where these have relevance, are appropriate, and are intelligible to the audiences.
- This implies greater awareness of the importance of statistical methods, and the cornerstone that 'effect size' analysis has in that process.
- This also implies that there is recognition of a large 'educational' training issue surrounding the use of more rigorous and complex analyses, especially as datasets extend over time.

Reference

FITZ-GIBBON, C.T. (1997). *The Value Added National Project: Final Report. Feasiblity Studies for a National System of Value Added Indicators*. London: SCCA.

8 Issues arising from the use of effect sizes in analysing and reporting research

Robert Coe

8.1 Introduction

This paper consists of three parts. The first argues the case for the use of effect size measures in analysing and reporting quantified data in educational research; the second explores some of the problems and complexities of doing just this; finally, the third part attempts to draw together some recommendations for appropriate use of effect sizes.

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. Where effect size is mentioned it is often as something of a footnote to a larger presentation of statistical significance testing. For these reasons it is well worth rehearsing the arguments for the use of effect size measures in reporting quantitative research.

Before launching into the argument, it is appropriate to clarify what is meant by effect size in this context. Although there are a number of alternative measures (outlined in this chapter), the argument here will focus mainly on the standardised mean difference, i.e. the difference between the mean values for two groups, divided by an estimate of the population standard deviation. It will also be assumed that, alongside any estimate of effect size, some indication of its margin of error, for example a 95% confidence interval, will be provided. Coe (2002) provides an introduction to these concepts.

In arguing the case for effect size, one should not lose sight of the fact that its use and interpretation can be problematic. In the heat of the argument for preferring an emphasis on effect size to the traditional hypothesis test, the limitations of the former are sometimes overlooked. It is important that these limitations should be made explicit and discussed if we are not to merely replace one set of unsatisfactory procedures with another. A starting point for this discussion forms the second part of this chapter, after the case for using effect size has been outlined.

8.2 The case for using effect size measures

Much of this argument will be familiar and has been presented many times, often in the context of criticisms of the use of statistical significance testing (e.g. Cohen 1969; Kirk, 1996, 2001; Harlow *et al.*, 1997; Thompson, 1999, 2002a; Wilkinson *et al.*, 1999). Here the argument focuses on the advantages of using effect size as an alternative, or supplement, to the use of hypothesis tests, drawing where appropriate on arguments about the deficiencies of exclusive use of the latter. Five broad arguments are presented, some of which contain multiple strands.

8.2.1 Effect size enables uncalibrated measures to be interpreted

Perhaps the most obvious motivation for the use of effect size is that it allows meaning to be given to a difference recorded by an unfamiliar instrument and reported on an unknown scale. By using the familiar concept of the standard deviation, it allows the difference to be calibrated in terms of the amount of variation within the overall population.

To illustrate this, consider a hypothetical example of a questionnaire on teachers' perceptions of their training needs. An overall scale has been created from the average of seven items in the questionnaire, each coded on a four-point Likert scale. Data from two comparisons are presented in Table 8.1.

Table 8.1 Comparison by age and sex of perceptions of training needs

Age 20–40			Age 41–65			Female			Male		
Mean	**N**	**SD**	**Mean**	**N**	**SD**	**Mean**	**N**	**SD**	**Mean**	**N**	**SD**
2.98	389	0.87	2.09	345	0.95	2.64	451	1.05	2.44	283	1.11

The table shows that younger teachers on average perceived a stronger need for training than their older colleagues and that females felt more need for training than males. Both differences are statistically significant ($p<0.05$), and in reporting such data it would be common to see no further comment on the size of the two differences. However, it is clear from the numbers that the difference between the two age groups is considerably bigger than that between the sexes. In fact, expressed as a fraction of the standard deviation of scores on the 'perception of training need' scale, the effect size of the former is 0.98, while that of the latter is 0.19 – about one fifth of the size. The use of effect size to calibrate these comparisons not only highlights their relative sizes, but also enables each to be interpreted. Here, for example, although there certainly is a difference between males and females, it is quite small. The difference between older and younger teachers, on the other hand, is substantial. Further discussion of the interpretation of effect sizes can be found in Coe (2002).

8.2.2 Effect size emphasises amounts, not just statistical significance

The arguments under this heading essentially relate to identified criticisms of the use of significance tests that are at least to some extent mitigated by use of effect size as an alternative. The argument comes in three parts.

Beyond dichotomies

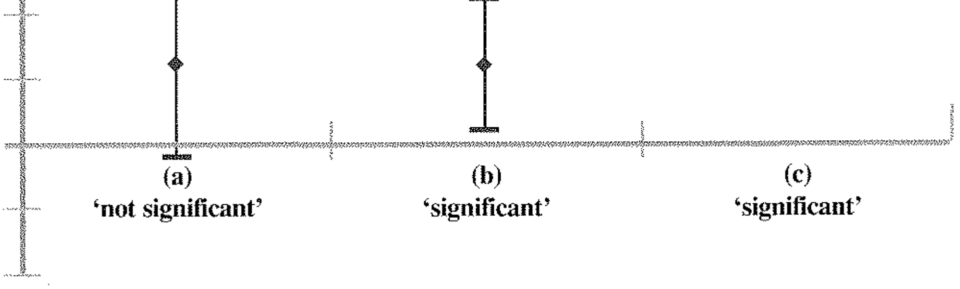
The dichotomous outcome of a significance test is often inappropriate in drawing inferences from data. In most research contexts (as opposed, say, to quality control) it is not appropriate to have to make an all or nothing decision about whether to accept or reject a particular null hypothesis. It is absurd to have to conclude one thing if the result of an experiment gives p = 0.051 and the exact opposite if it were 0.049. (Oakes, 1986). The true/false dichotomy can easily be transcended by the use of an effect size estimate with a confidence interval.

Amount is important

One of the most telling criticisms of the use of significance tests is that they leave out the most important information: the size of the effect. It is not enough to know, as Tukey (1969) has said, 'if you pull on it, it gets longer'. Scientific advance requires an understanding of 'how much'. Significance tests do not tell us how big the difference was, or how strongly related two variables were. Instead, they say more about how large our sample was (Thompson, 1992). A great deal more information can be extracted from an experiment if the focus is on parameter estimation, rather than hypothesis testing (Simon, 1974).

An illustration of the importance of amounts rather than just directions can be seen in Figure 8.1, which shows three comparisons from three hypothetical independent experiments, illustrated graphically as effect size estimates and their confidence intervals. In (a), a significance test would show the effect to be not significant, while for (b) and (c) it would be statistically significant. However, the presentation of effect sizes shows clearly that the results of (a) and (b) are actually the same; differences in sample size or design prevent the former reaching statistical significance. On the other hand, the result in (c) is quite different and not consistent with the other two.

Figure 8.1 The failure of significance tests to quantify the size of a difference



The meaning of significance

There is room for debate about precisely how the 'significance' of a result should be defined. Issues such as its policy or theoretical implications, costs, benefits and feasibility, together with the strength of evidence underpinning it, should probably all be considered (Thompson, 2002b; Leech and Onwuegbuzie, 2003). However, what is clear is that significance tests are widely presented and interpreted as conveying the size of the effect and its replicability (Oakes, 1986), but that in fact they do neither. Certainly, for results with practical or policy applications, the effect size is arguably a better index of significance than a significance test.

8.2.3 Effect size draws attention to the margin of error

Statistical power

By considering the power of significance tests reported in social science journals, Cohen and others (see Cohen, 1990) have shown that the majority of studies published have a less than even chance of rejecting the null hypothesis, even where there is in fact a medium-sized effect. In other words, failure to reject the null hypothesis typically tells you absolutely nothing, other than that your sample was probably too small. What is extraordinary is that in the time since Cohen (1969) originally published this finding, the situation seems not to have improved (Cohen, 1990), suggesting that concern with statistical power is not paramount in designing research (but see below). The use of an effect size and confidence interval, represented as in Figure 8.1, makes the margin of error around a result very apparent and it is hard to imagine that such inappropriate use of significance tests could have continued had this kind of representation been more widespread.

It may also be noted here that in order to calculate the likely power of a comparison one has to make some kind of assumption about the size of the difference in a population. Such calculation is greatly simplified if the difference is expressed in terms of effect size, and indeed this was Cohen's (1969) original reason for introducing the concept. One could therefore say that as well as drawing attention to the issue of power, the use of effect sizes is actually a requirement for its calculation.

Synthesis rather than disagreement

One interesting comment on the adverse effect of significance testing is that a good many disagreements in social science are simply due to sampling variation (Hunter and Schmidt, 1996). Many apparently different findings are in fact perfectly consistent with each other, within the margins of statistical error. Significance testing greatly exaggerates these differences, stressing individual results at the expense of an integrated overview of all the available evidence, together with its associated uncertainty. It could therefore be argued that the use of effect sizes might help to reduce this adversarial tradition in social science, in favour of a more consensual, synthetic approach.

8.2.4 Effect size may help to reduce reporting bias

The file drawer problem

The 'file drawer problem' (Rosenthal, 1979) refers to the over-representation in published work of statistically significant results, leading to overall bias. Research syntheses based on easily available studies are liable to over-estimate the size of an effect, because those that failed to achieve statistically significant results are less likely to be published. It is certainly possible that increased use of effect size would reduce this bias.

Within-study reporting bias

Even within a study it is impossible to know how many 'non-significant' relationships have been tested, consciously or not, in order to find the 'significant' ones that are presented. The statistical significance of a result depends not just on the data, but on the way such findings were sought. This is a particular problem when blanket, multiple significance tests are used to identify 'significant' results (Wilkinson *et al.*, 1999).

8.2.5 Effect size allows the accumulation of knowledge

Evidence from different studies can easily be combined

One of the most obvious 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' (Glass *et al.*, 1981).

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 (see 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 recognition that scientific advancement proceeds by the accumulation of knowledge, not by results considered in isolation, still seems to be a long way from being accepted by all educational researchers.

Small studies count

An important consequence of the capacity of meta-analysis to combine results is that even small studies can make a significant contribution to knowledge. For example, 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. Even Fisher, who is often credited with much of the responsibility for the evils of significance testing, regarded the 5% level as arbitrary and took as a basis for knowledge the repeated finding of results at this level, rather than any single highly 'significant' result (Tukey, 1969). However, because of the orthodoxy of significance testing, these small studies may never be done, having been rejected at the planning stage as having insufficient power. 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). A large number of studies with small samples and similar results may provide more evidence about a phenomenon than a single large study, but taken individually none of them may have the power to achieve statistical significance.

8.3 Problems in using effect size measures

It has been argued above that the wider use of effect sizes has the potential to reduce some of the anomalies that have been identified with inappropriate uses of statistical significance tests. However, it is important to remember that more careful and appropriate use of significance tests could also have the same effect. Moreover, there is no guarantee that the unthinking use of significance tests would not simply be replaced by an unthinking use of effect sizes, without any significant improvement in research practice (Wainer and Robinson, 2003).

Much of the debate on the problems of significance testing has been characterised by strong feelings and polarised positions, making it hard to see the true complexity of the situation. It is important in advocating the use of effect sizes that their benefits are not over-claimed and their limitations not overlooked. The following section outlines some of these limitations and complexities, pointing out some problems that may arise from their use.

8.3.1 Which 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 briefly 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 the appropriate statistic is usually called 'eta-squared', η^2).

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'.

When is an effect not 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

One problem with the use of the 'standardised mean difference' measure of effect size is that its interpretation is very sensitive to violations of the assumption of normality. For this reason, a number of more robust (non-parametric) alternatives have been suggested. Some examples are discussed by Coe and Merino (2003, pp. 171–2).

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) and others in this volume. 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).

Differences between proportions

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).

In fact, in the case where a single comparison is made between two proportions, one could argue that those proportions can themselves already be easily interpreted and that a complex statistical reformulation of them is of little value. The main value of the odds ratio comes in combining the results from different studies in meta-analysis.

Unstandardised (raw) difference

Just as one does not need to overcomplicate the interpretation of the difference between two proportions, there are some cases where a continuous variable is measured on a familiar scale and one can simply report the difference between the two means, without standardising. Examples of universally familiar scales would include outcomes involving time (measured in seconds, days, years, etc), money (pounds, dollars), length (metres) etc. Some scales could be assumed to be familiar within a particular context, for example A level grades in England.

8.3.2 Which standard deviation?

A pooled estimate

The standard deviation, used to standardise the standardised mean difference, should ideally refer to the whole population. One might expect that the control group would 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, since it is typically only half the size. Moreover, in studies where there is not a true control group 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 (Hedges and Olkin, 1985). The pooled estimate is essentially an average of the standard deviations of the experimental and control groups, otherwise known as the 'within-groups' standard deviation. The implications of choices about which standard deviation to use are discussed by Olejnik and Algina (2000).

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.

Statistical controls: residual standard deviation

An easy way to make an effect size substantially bigger is to divide the difference by the residual standard deviation, after introducing appropriate explanatory variables into a statistical model, instead of just the standard deviation of the raw outcome. For example, in a regression (or multilevel) model in which half the variance is accounted for ($R^2 = 0.5$), an effect size calculated using residual standard deviation would be 40% larger than it would have been if the standard deviation of the raw outcome had been used. For this reason, one should certainly at least make it clear which kind of effect size is being used, and present the variance accounted for by the model (R^2). Further, the use of residual rather than raw standard deviation must require some justification.

It is possible to think of some circumstances in which this kind of inflation may be justified. If the residuals relate to the gain between a pre-test and post-test score on the same measure, then they could arguably be interpreted as a direct measure of the progress made between the two measurements. In this case, one could perhaps talk about the effect on progress of an intervention (as compared to the effect on the raw outcome post-test scores) and compute an effect size as the difference between the mean residual gains of the two groups, divided by the residual standard deviation. The interpretation of such an effect size, however, and consequently its use in any comparison or synthesis with other effect size estimates, would be very problematic, since it would be highly sensitive to the strength of the correlation between pre- and post-test measures. An example of the kind of anomaly this could generate is that effect sizes from short-term interventions would be likely to be larger than for those that were more sustained, since the shorter test-re-test interval could be expected to result in higher correlations.

Populations with restricted range

Another problem that relates to the standard deviation used to calculate an effect size arises when the samples used are drawn from a group with a more restricted range than some theoretical overall population. 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.

8.3.3 Measurement reliability

A further 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 random '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, an experiment could be conducted where an experimental and control group were each given two different post-tests measuring the same construct, one more reliable than the other. This would typically be the case if, say, one test were longer than the other, or consisted of more discriminating items, or was better targeted at the particular range of scores found in the two groups. For whatever reason, though, the less reliable test would contain more error in its scores and would therefore have a larger standard deviation. 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).

8.3.4 Non-normal distributions

The interpretations of effect-sizes in terms of percentiles at which they overlap depend on the assumption that both control and experimental groups have a normal distribution. Needless to say, if this assumption is not true then the interpretation may be altered, and in particular, it may be difficult to make a fair comparison between an effect-size based on normal distributions and one based on non-normal distributions.

An illustration of this is given in Figure 8.2, 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 8.3. 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). This is quite a substantial difference and illustrates the problem of interpreting effect sizes when the distribution is not known to be normal.

Figure 8.2 Comparison of Normal and non-Normal distributions

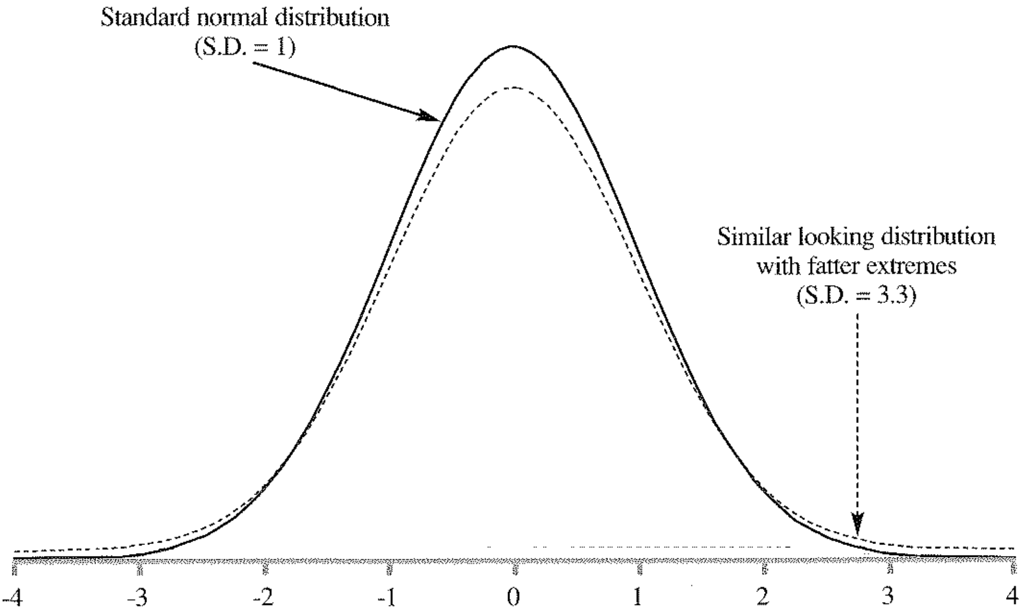
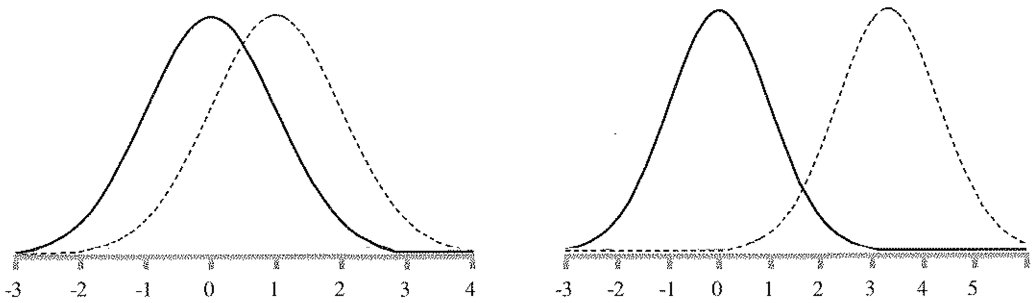


Figure 8.3 Normal and non-Normal distributions with effect-size = 1



(a)

(b)

8.3.5 'Small', 'medium' and 'large' effects

In his original presentation of the standardised mean difference, Cohen (1969) cautiously provided the interpretation that within psychological research, 0.2 could be considered a small effect, 0.5 medium and 0.8 large. Unfortunately, and particularly in text books where effect size is given a brief and somewhat token mention, these interpretations seem to have become part of the orthodoxy of effect size interpretation. This is unfortunate because the issue of whether an effect should be considered 'large' depends on a number of factors. These might include the costs of implementing the intervention, its practical feasibility, the benefits associated with the difference produced and the value attached to those benefits, as well as the sizes of other effects produced by comparable interventions in the same context and with the same outcome (Glass *et al.*, 1981, p. 104). The interpretation of the size of an effect should also depend on the technical issues outlined above, such as measurement reliability and range restriction. Coe (2002) gives examples of effect sizes from a range of interventions in education and elsewhere along with some further interpretations of effect sizes in order to provide a context for these kinds of comparisons.

8.3.6 Incommensurability

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; when they are both effect sizes this temptation is particularly seductive. 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 is very easy, once standardised, scale-free 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 above), 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 in the fidelity of its delivery, or that the same treatment may have had different levels of intensity in different studies. This applies equally to the 'treatment' applied to the control group; a comparison based on a 'treat as normal' control may be quite different from a 'withhold treatment' group. In any of these cases, it makes no sense to average out their effects.

8.4 Summary and recommendations

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 a number of assumptions, including 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.

The 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. Moreover, it has been argued that the wider use of effect sizes could help to avoid some of the problems that have been associated with significance tests, including their requirements for inappropriate dichotomous decisions and their tendency to ignore any information about the magnitude of a difference and lead to a reclaiming of the meaning of the word 'significance' to be more in line with common usage. Effect size use may also draw attention to the vital issue of statistical power in making comparisons and reduce the potential both for sampling variation to be misinterpreted as true difference and for the failure to find a clear difference to be misinterpreted as evidence of no difference. This in turn may help to reduce reporting bias, both across and within studies. Effect sizes enable the results from different studies to be combined, leading to a more rational approach to the accumulation of knowledge and freeing research from the restrictive requirement that an individual study must single-handedly and in isolation provide definitive answers to its research questions.

Despite these potential advantages, however, the use and interpretation of effect sizes is not without its problems. Choices about which standard deviation to use for calibrating differences as well as any restriction of range, measurement reliability or deviations from normality can all spuriously and covertly affect the interpretation of a calculated effect size. It is also very easy to combine or compare effect sizes that are essentially incommensurable, so creating a seductively simple but meaningless statistic.

It is therefore recommended that researchers should:

- calculate and report standardised effect sizes, with confidence interval or standard error, for all comparisons where a statistical significance test might have been done
- show these effect sizes and their confidence intervals graphically
- report all relevant comparisons regardless of whether confidence intervals include zero (i.e. whether they are statistically significant), especially if the comparison was planned before any data were seen
- interpret effect sizes by comparison with known effects and in relation to familiar metrics
- report un-standardised raw differences whenever the outcome is measured on a familiar scale
- interpret the significance of an effect with regard to issues such as its:
  - effect size
  - theoretical importance
  - associated benefits
  - associated costs
  - policy relevance
  - feasibility
  - comparison with available alternatives
  - sampling error (statistical 'significance')
- not use the word 'effect' (with or without 'size') unless a causal claim is intended and can be justified. Be cautious about the calculation and

interpretation of standardised effect sizes whenever:

- sample has restricted range
- population is not known to be normal
- outcome measure has low or unknown reliability
- outcomes have been statistically adjusted (residuals)

- always report reliability of measures, extent of restriction, correlations (or R^2) in these cases
- encourage small studies with low power and statistically non-significant effects still to be conducted, reported and published, provided they are free from bias
- synthesise the results of compatible studies using meta-analysis
- beware of combining or comparing effect sizes from studies with incommensurable outcomes, different operationalisations of the same outcome, different treatments, or levels of the same treatment (including control group 'treatments'), or measures derived from different populations.

References

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

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, 254–63.

COE, R. (2002). 'It's the effect size, stupid: what effect size is and why it is important.' Paper presented at the British Educational Research Association Annual Conference, Exeter, 12–14 September [online]. Available: http://www.leeds.ac.uk/educol/documents/00002182.htm [4 March, 2004].

COE, R. and MERINO, C. (2003). '¿Magnitud del Efecto?: Guía inicial para usuarios', *Revista de Psicología*, **21**, 1, 146–77.

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

COHEN, J. (1990). 'Things I have learned (so far)', *American Psychologist*, **45**, 12, 1304–12.

FITZ-GIBBON, C.T. (2002). 'A typology of indicators for an evaluation-feedback approach.' In: VISSCHER, A. and COE, R. (Eds) *School Improvement Through Performance Feedback*. Lisse: Swets and Zeitlinger.

FLEISS, J.L. (1994). 'Measures of effect size for categorical data.' In: COOPER, H. and HEDGES, L.V. (Eds) *The Handbook of Research Synthesis*. New York, NY: Russell Sage Foundation.

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, NY: Academic Press.

HUBERTY, C.J. (2002) 'A history of effect size indices', *Educational and Psychological Measurement*, **62**, 2, 227–40.

HUNTER, J.E. and SCHMIDT, F.L. (1996). 'Cumulative research knowledge and social policy formulation: the critical role of meta-analysis', *Psychology, Public Policy and Law*, **2**, 2, 324–47.

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, 350–86.

KIRK, R.E. (1996). 'Practical significance: a concept whose time has come', *Educational and Psychological Measurement*, **56**, 5, 746–59.

KIRK, R.E. (2001). 'Promoting good statistical processes: some suggestions', *Educational and Psychological Measurement*, **61**, 2, 213–18.

LEECH, N.L. and ONWUEGBUZIE, A.J. (2003). 'A proposed fourth measure of significance: the role of economic significance in educational research.' Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, 21–25 April.

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, 669–76.

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

OLEJNIK, S. and ALGINA, J. (2000). 'Measures of effect size for comparative studies: applications, interpretations and limitations', *Contemporary Educational Psychology*, **25**, 241–86.

ROSENTHAL, R. (1979). 'The "file drawer problem" and tolerance for null results', *Psychological Bulletin*, 86, 638–41.

ROSENTHAL, R. (1994) 'Parametric measures of effect size.' In: COOPER, H. and HEDGES, L.V. (Eds) *The Handbook of Research Synthesis*. New York, NY: Russell Sage Foundation.

RUBIN, D.B. (1992). 'Meta-analysis: literature synthesis or effect-size surface estimation', *Journal of Educational Statistics*, **17**, 4, 363–74.

SIMON, H.A. (1974). 'How big is a chunk?' *Science*, **183**, 482–8.

SNYDER, P. and LAWSON, S. (1993). 'Evaluating results using corrected and uncorrected effect size estimates', *Journal of Experimental Education*, **61**, 4, 334–49.

THOMPSON, B. (1992). 'Two and one-half decades of leadership in measurement and evaluation', *Journal of Counseling and Development*, **70**, 434–38.

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 [online]. Available: http://acs.tamu.edu/~bbt6147/aeraad99.htm

THOMPSON, B. (2002a). 'What future quantitative social science research could look like: confidence intervals for effect sizes', *Educational Researcher*, **31**, 3, 25–32.

THOMPSON, B. (2002b). '"Statistical", "practical" and "clinical": how many kinds of significance to counselors need to consider?' *Journal of Counseling and Development*, **80**, 64–71.

TUKEY, J.W. (1969). 'Analyzing data: sanctification or detective work?' *American Psychologist*, **24**, 2, 83–91.

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.

WAINER, H. and ROBINSON, D.H. (2003). 'Shaping up the practice of null hypothesis significance testing', *Educational Researcher*, **32**, 7, 22–30.

WILCOX, R.R. (1998). 'How many discoveries have been lost by ignoring modern statistical methods?' *American Psychologist*, **53**, 3, 300–14.

WILKINSON, L. and TASK FORCE ON STATISTICAL INFERENCE (1999). 'Statistical methods in psychology journals: guidelines and explanations', *American Psychologist*, **54**, 8, 594–604.

9 Effect size: a statistician's pseudo-concept?

Ray Godfrey

This paper uses a philosophical approach to raise questions about the usefulness or otherwise of the notion of effect size.

First a word about how philosophy works. For this paper I wish to make a point about the use of effect sizes by employing the distinction between a concept and a pseudo-concept. The distinction is something I have invented for the purpose of making the point about effect sizes. As with any distinction, once I have explained what I mean by pseudo-concept it will not be difficult for others to argue that everything is a pseudo-concept or that nothing is a pseudo-concept or that the distinction is totally unintelligible. It would be disappointing if my idea aroused so little interest that nobody bothered to do this. What I hope to achieve is to make my point before readers become disenchanted. Once the point is made, the distinction becomes unimportant. If I am very successful, readers will find my arguments so convincing that they will not be able to believe that my point is not obvious, so obvious that it could have been stated much more briefly and possibly even so obvious that it did not need stating at all. If I am unsuccessful, the paper will be dismissed as an attack on advanced statistics.

Rather like statistics, theories of meaning became much more sophisticated during the twentieth century. Previously, people had regarded a word as having a meaning and the meaning of a sentence or proposition as being formed by combining the meanings of all the words in it. This view was often combined with a belief that the sentence or proposition was true if the picture it painted of the world corresponded to the way the world actually was. 'The cat sat on the mat' is true precisely if the bit of the world called 'the cat' is connected to the bit of the world called 'the mat' in exactly the way that the phrase 'sat on' suggests.

The later work of Wittgenstein employs a much more subtle (and perhaps more productive) approach of looking for meaning in the way words are used in the language games of life. He is able to make some quite profound points that would not be easily accessible in a more formal and restrictive philosophy (see Wittgenstein (1969) for some useful insights into the nature of knowledge). 'The cat sat on the mat' is true if used in circumstances where the rules of the language game make this an appropriate thing to say.

Some later philosophers seem to have stopped trying to link meaning to the world or to life at all.

As in statistics there are some things that are best done using a simple t-test, so I think there are some points best made using unsophisticated theories of meaning. Take a concept to be something that has sense and possibly reference. The concept of 'cat' has a sense that we have to be aware of if we are to use the word in the same way as other English speakers. It also has a reference: it refers to each and every cat in the world. The sense is related closely to the reference (see Frege (1952) for a thorough treatment of this). Here we have a word that has meaning.

Some words cannot be discussed so easily in terms of sense and reference. It is easier to follow the later Wittgenstein and look instead at their use. These words operate in the grammar of a language as if they were concept-related words, but actually carry no sense. This does not mean that they are nonsensical. It does mean that when you try to explain their meaning the task turns out to be not only difficult but impossible. I shall call these pseudo-concepts.

One way in which pseudo-concepts come about is by illusions created by grammar. In modern English it is very easy to generate nouns simply as turns of phrase. If we see someone who behaves intelligently, it is no more than correct use of grammar to say that he is intelligent or that he shows intelligence. We can then think about intelligence as if it were a thing, some sort of entity that the person possesses. There is no entity to which 'intelligence' refers nor any sense by which it might refer to something (see Ryle (1949) and Chomsky (1957) for similar thoughts). This is a pseudo-concept. Psychology would be very difficult without such pseudo-concepts. The way in which psychologists (and others) employ them is all that there is to discuss if we wish to see their meaning. But if psychologists reach conclusions that cannot be translated back into terms of people behaving intelligently, something very strange and possibly dangerous has occurred.

The point I am making here is related to the view of infinity and infinitesimals taken by some philosophers of mathematics (see Boyer

(1959, pp. 217–223) for a summary of Leibniz's position). Infinitesimals are taken as meaningless ideal elements in mathematics, which can be used in proofs only provided that logical rules are adhered to and only so long as the conclusions are expressed in terms of concepts that have meaning. One way to look at the constructivist school of mathematical philosophy (not to be confused with the constructivist school of educational psychology and philosophy) is as an attempt to avoid the use of pseudo-concepts in order to avoid being led into error by them.

Pseudo-concepts are not always (or even usually) a bad thing. Advanced mathematics could not be engaged in without them. Even some quite simple mathematics only becomes accessible once students develop pseudo-concepts. One type of pseudo-concept that is relevant here is the procept, an idea introduced by David Tall (1991) (see also Grey and Tall, 1994). The point is that when students start to work with something like the fraction 3/5 it is initially shorthand for a process that you go through. When you see '3/5', you have to split something into five parts and use only three of them. As long as the student sees this as a process they can make no progress with fractions. It is necessary to become so familiar with '3/5' that you think of it as if it were an entity in itself. Until 3/5 becomes a thing in its own right it is difficult to make sense of '3/5 + 2/3' and absolutely impossible to make sense of things like '$\ln(\pi^{3/5})$' . The best way to explain to someone the meaning of 3/5 in '$\ln(\pi^{3/5})$' is to show them how it is used.

Another example is directed numbers. As long as you see '–2' as something to do with counting along a number line or something related to how much money you have in your bank balance it will not be possible to make much sense of ' $-2 \times -3 = +6$'.

There are two types of danger related to pseudo-concepts that I think are relevant to the debate over effect sizes. The first of these concerns learning to do mathematics or at least trying to understand it. The basis for my comments is thirty-three years of reflection on attempts to teach mathematics or statistics to people of all ages and of varied educational attainment. Part of this reflection has involved reading research reports, but my views do not depend upon research evidence.

One of the things that can make it difficult to understand mathematics is a failure to realise that the ideas involved are not concepts but pseudo-concepts. In a sense they have uses but no meanings. Learners can feel that they do not *understand* a topic because, although they can carry out all the required manipulations, they cannot see what it *means*. At the Open University M101 Summer School there used to be a session on relations. Students would come to the Summer School having covered the unit on relations and in some cases having correctly answered all the assessment questions. They would attend the session in a state of bafflement. The best strategy for dispelling this bafflement seemed to be to persuade them that they already understood all there was to understand and there was no mysterious entity called a relation that they needed discover. 'Relation' is just a word that we use in situations where you can do all the things they had been doing. Many people went away much happier once they stopped trying to understand the meaning of it. After passing that hurdle it was possible to speak and think as if the word did refer to something, a mathematical object.

The misunderstandings of percentages caused by people thinking that a 'per cent' is a thing that can be combined with other 'per cents' in pretty much the same way as other things are too common to need an account here.

Statistics provides a further basic example. One of the first notions that many students of statistics encounter is that of standard deviation. I suspect this is responsible for much distaste for statistics as a whole. First, teachers often insist that students calculate standard deviations. This is probably one of the most difficult calculations the students have ever been asked to carry out and could well put them off. Alternatively, they may be allowed to obtain standard deviations using calculators or computers. In either case many of them are dissatisfied because they do not know what a standard deviation *is*. They want to know what it *means*. In response to this the teacher can recapitulate the method of calculation, which occasionally satisfies some, or give examples of the way in which it is used. You can use standard deviations to compare the dispersion of different samples, to carry out t-tests or even to calculate effect sizes, but none of this tells you what they mean. Students will not make much progress in statistics until they can accept that a standard deviation is just a number that is calculated in such-and-such a way, is used in such-and-such a way and has no independent meaning in between times. Many students who later go on to be successful do this without effort and without thinking. Many others find it a stumbling block.

The notion of an effect is more troublesome than that of standard deviation. A statistical effect often has nothing to do with cause and effect, thus causing some confusion. More importantly there is no such thing as *the* effect, pure and simple, of gender (say) on key stage 2 attainment. The effect of gender depends on the statistical model in question. Even within the same type of model, the inclusion of different combinations of other explanatory variables alongside gender may leave gender with more or with less effect. Robert Coe's criticisms of the notion of 'effect' are also relevant here. The word is suggestive of causality. It is also suggestive of reality.

Turning to 'effect size', this suffers from all the problems associated with 'standard deviation', all those associated with 'effect' and a few more of its own.

Effect size is partly a matter of experimental design. Any book on the subject will tell you how to calculate the sample size necessary in order to achieve a given power for a particular effect size and for a given critical significance level. A good book (Lipsey (1998) for example) will tell you that increasing the sample size may be expensive or even impossible and that you should first ensure that your experiment is designed optimally so that the effect size is as large as possible. Use more refined measuring instruments. Control the experimental conditions more finely. Reduce diversity in the participant groups. Make the difference between experimental and control conditions more marked. These and other devices ensure that against the background of the same real-world phenomena the effect size is increased, so for the same sample size the power will be increased. It matters little here whether effect size is a concept or a pseudo-concept. It fulfils a vital function in the thinking of the experimental designer and no one else need ever know about it. Most people only wish to know the sample size necessary to achieve their aims.

This book is concerned with other, and to my mind more questionable, aspect of effect sizes: their use in communicating results.

Very few people understand the notions of statistical significance, power and the like. Even the first edition of Cohen's classic work on power and effect sizes contains one blunder in this respect (Cohen (1988, p. 6) corrects 'virtually no power' to 'no power, since that conclusion is inadmissible'). Furthermore, statistical significance is not the same as educational importance. A poorly designed study may well fail to find significant evidence for an important educational effect. Neither is this the same as finding a number in our statistical results which looks a lot bigger than other numbers appearing in the same context. For example in regression models the constant term may be quite large compared with other parameters, but be of no educational interest and not necessarily significant. For these and other reasons it is a good idea to present results in a way which appears to answer the questions: 'Exactly what difference does this variable make?', 'Exactly what difference does class size make to attainment?', 'Exactly what difference does ethnicity make to school exclusions?' They want to know, in Cohen's phrase, 'the degree to which the phenomenon exists' (Cohen, 1988, p. 4). The answer people want to these questions is something that has both sense and reference in the real world. I suggest that, in effect size, what we offer them is a pseudo-concept.

Effect size is sometimes used to refer to something that does have meaning in the world beyond statistics. Effect size may be measured in real units. In medicine there is a reasonable chance that virtually all studies on a particular topic are concerned with measuring the same phenomena on the same scale, and giving effect sizes in terms of blood pressure or cholesterol levels appears unproblematic. There is less chance of this occurring in education. There may be a number of studies using the same NFER standardised test, there may be a number of studies using A-level point scores, but there is no reason why everybody interested in a topic should consider these measures to be the best for their purposes, especially if their work is not confined to England. Some of my work is concerned with absence from school and you might think that an absence rate is an absence rate, but the DfES Performance Tables give four different measures and there is no guarantee that any one of them, or any particular combination of them, will be best suited to every study of absenteeism.

Effect sizes can also be given in absolute units. These may be difficult to understand but they do at least give an unambiguous measure of how an intervention has changed things. They are frequently related to binary outcomes. In medicine, mortality or morbidity rates are usually unambiguous (though not necessarily accurately measured) and changes in these can be given in terms of odds ratios or a range of other absolute measures. In education binary outcomes are sometimes important but frequently of only secondary interest. Much of my work is concerned with exclusions from school, but in many ways the least interesting thing you can ask about a pupil's school history is whether they have been permanently excluded from a school in a formal way that leads to an entry in the official exclusion statistics. An intervention which could be evaluated simply in terms its odds ratio for a pupil's chance of official exclusion would probably not be a very interesting one.

These two ways of conveying effect size can lull people into thinking that the phrase 'effect size' in itself has some clear meaning. They can encourage people to look at standardised effect sizes, such as Cohen's d as if they were real units or as if they were absolute measures. Neither is the case.

Brown *et al.* (2003) refer to a single effect size in their brief summary of a large research project. They state that the implementation of the National Numeracy Strategy 'demonstrated an effect size of 0.18' on children's attainment. Schweinhart *et al.* (1993) make very full use of standardised effect sizes in the many tables of their report on the effects of the Perry High/Scope project, though their argument is developed largely in terms of significance. To say intervention has a small effect size is not the same as saying that it has no educationally important effect.

Standard effect sizes do not tell us about a two-sided relationship between the intervention and the real world. They tell us about a three-sided relationship between the intervention, the real world and the study design. To calculate Cohen's d what you do is very much like carrying out a t-test but stopping just before you take sample size into account and well before you calculate a significance level. Given the role of d in calculating samples sizes necessary to achieve given significance levels and required power level, this is not surprising. If for a given d and an intended sample size the power is too low, there are two approaches to increasing the power. One is to increase sample size. The other is to redesign the experiment to reduce error variance. A retrospectively calculated Cohen's d tells us something about the relationship between the effect of the intervention on the real world and the quality of experimental design.

Cohen's d is not an absolute measure, though it is (like all ratios) dimensionless. The impact of the intervention is given not in absolute terms, but in relation to the standard deviation of the variable used to measure the effect. Standard deviation itself suffers from the drawbacks associated with pseudo-concepts. This makes it difficult to build upon it anything other than further pseudo-concepts. But the choice of standard deviations raises even more problems for anyone wishing to claim that 'effect size' has a clear meaning. Debates over whether to estimate population standard deviations from the control group or from pooled data are relatively unimportant. What removes any residual chance of finding a clear meaning in Cohen's d is the question of which population the standard deviation is calculated or estimated for. With a standardised test, the NFER will give you a population standard deviation, but this will not be the same as the standard deviation for the stratum of pupils at whom a particular intervention is aimed. Nor is it necessarily true that the standard deviation found in the mathematically poor achievers in three study schools is the same as the standard deviation in the group of mathematically poor achievers that concern teachers in another school who are trying to judge the effectiveness of the intervention.

Complex procedures for estimating 'real' effect sizes in the face of abstruse statistical difficulties or for calculating effect sizes for forms of statistical analysis further and further removed form the simple t-test exemplified in Peter Tymms' paper (Chapter 5) are very useful for statisticians but do nothing to dispel the idea that effect size is a meaningful thing that refers to some observable characteristic of the real world beyond statistics that statisticians are anxiously trying to grasp and express. We should at least try to find forms of expression that convey to the broader public the nebulosity of the ideas involved and the intimate dependence upon the nature of effect size and the nature of the statistical design and analysis that lie behind it. Alternatively we could follow the lead of Ian Schagen (in Chapter 3) and look for other simpler ways of presenting the results of complex models, models that may well have been guided in their development by considerations of effect sizes.

If there are absolute units or generally understood units in which effect sizes can be given and mean effect sizes calculated, all well and good. But if effect sizes can only be given in standardised form or if aggregation of evidence can only be done using standardised effect sizes, then we have to accept that the notion we are using is a pseudo-concept, extremely useful within the world of statisticians but of no real meaning in the world beyond that.

References

BOYER, C. (1959). *The History of Calculus and its Conceptual Development*. New York, NY: Dover.

BROWN, M., ASKEW, M. and MILLET, A. (2003). 'How has the National Numeracy Strategy affected attainment and teaching in Year 4?' *Proceedings of the British Society for Research into Learning Mathematics*, **23**, 2, 13–18.

CHOMSKY, N. (1957). *Syntactic Structures*. The Hague: Mouton.

COHEN, J. (1988). *Statistical Power Analysis for the Behavioural Sciences*. Second edn. London: Lawrence Erlbaum.

FREGE, G. (1952). 'On sense and reference.' In: GEACH, P. and BLACK, M. (Eds) *Translations for the Philosophical Writings of Gottlob Frege*. Oxford: Basil Blackwell.

GRAY, E. and TALL, M. (1994). 'Duality, ambiguity, and flexibility: a proceptual view of simple arithmetic', *Journal for Research in Mathematics Education*, **25**, 2, 116–40.

LIPSEY, M. (1998). 'Design sensitivity: statistical power for applied experimental research.' In: BICKMAN, L. and ROG, D. (Eds) *Handbook of Applied Social Research Methods*. London: Sage

RYLE, G. (1949). *The Concept of Mind*. Harmondsworth: Penguin.

SCHWEINHART, L., BARNES, H. and WEIKART, D. (1993). *Significant Benefits: the High/Scope Perry Preschool Study Through Age 27*. Ypsilanti, MI: High/Scope Press.

TALL, D. (Ed) 1991). *Advanced Mathematical Thinking*. Kluwer Academic Publishers.

WITTGENSTEIN, L. (1969). *On Certainty*. Oxford: Basil Blackwell.

10 Effect size: what does it mean?

Caroline Sharp

When my colleague Ian Schagen approached me to be a discussant for this seminar, my immediate reaction was: 'But I don't know anything about statistics'. Ian replied: 'I know, that's why I want you to do it.' I think what he was trying to say was that in approaching the issue of effect size, the organisers of this seminar wanted to address the concerns and interests not just of statisticians but also of their non-statistician colleagues and clients.

My approach to the task has been to ask whether and to what extent effect size is useful in educational research. Does it convey information that is additional to that offered by other types of statistical analyses? If so, what kind of information and how is this best interpreted? What are its limitations and what are the potential pitfalls that researchers should be aware of when seeking to use effect size to make inferences, draw conclusions or make recommendations?

I am grateful to Robert and Ray for addressing these questions in Chapters 8 and 9. I understand that their positions differ: Ray's paper is much less enthusiastic about the use of effect size than is Robert's. But let me begin with what they have in common.

Both writers are concerned with the way in which statistics convey meaning to non-statisticians. Let's start with the name. As a sometime editor of my colleagues' draft reports, I have enough trouble with the term 'significance'. Because significance has both an everyday meaning (something like 'important' or 'noteworthy') and a more technical meaning of 'statistical significance', this can lead to confusion. Are we in danger of suggesting too much in the very name 'effect size'? As Robert's paper points out, 'effect' carries an implication of 'cause and effect' and thus may mislead people to believe that a causal relationship has been established, whether or not this is the case.

Robert therefore suggests that we should use the term conditionally, only when a claim for causality is intended. This is an interesting suggestion, but by doing this we need to be clear that the term is no longer being attached to a particular concept, but rather to the intended interpretation of the concept in a specific context. Unless we can come up with an alternative way of naming the 'effect size' calculation, I can imagine that this might cause difficulties for researchers trying to communicate with statisticians (could you do that calculation for me, you know the one that's called an 'effect size' some of the time…).

This brings us on to the main thrust of Ray's paper: the pseudo-concept and its associated problems. What I understand Ray to be arguing is that effect size is a pseudo-concept because it has no real world existence or correspondence. Other pseudo-concepts mentioned in Ray's paper are intelligence, fractions, percentage and standard deviation.

Ray points out that pseudo-concepts, although useful to mathematicians, already cause much confusion in the real world because they have no concrete reality. Nevertheless, I would argue that such pseudo-concepts are indispensable in statistics, if not in life. I found Ray's discussion helpful in explaining the root causes of the difficulties we all encounter in understanding and applying such concepts. What I couldn't work out was what Ray was suggesting we should do about this in respect of effect size. Should we reject effect size because it is likely to cause confusion and lead to inappropriate decision making, or should we proceed with considerable caution?

In talking about the somewhat narrow issue of the application and interpretation of effect size, both Ray and Robert refer to one of the biggest challenges for applied social researchers: how do we provide information of use to policy makers and practitioners? The problem is that policy makers and practitioners often want simple answers, and get frustrated with researchers trying to over-complicate things. On the other hand, researchers and statisticians get frustrated when they can see the complexity of the issues involved, but find that policy makers and practitioners want simple messages: what does this mean and what should I do about it? This problem will not go away.

Like Robert, I can see a potential utility in calculating effect size. As we know, measures of statistical significance are only helpful to a point. Rather than concentrating on the boundary between a 'significant' and a non-significant relationship, effect size adds an appreciation of the strength of the relationship. It also provides a measure that enables comparisons to be carried out between independent studies.

I well remember a civil servant contacting the NFER urgently, saying that the minister was about to make a speech and wanted to know how much of an effect had been produced by the intervention we were evaluating and how well it compared to other educational initiatives, both in the UK and worldwide. Now, as it happened we had calculated effect sizes as well as using other statistical measures and we were able to make a calculation in relation to months of progress. Our problem in comparing initiatives was that very few other evaluation studies had reported effect size (or even standard deviations) so we simply couldn't make the comparisons the minister was seeking.

In conclusion then, what I take from these two papers is very useful food for thought. Ray has exposed me to the concept of the pseudo-concept (is that a *pseudo* pseudo-concept?) and has led me to reflect on its implications for translation of abstract concepts into something meaningful in the real world. Robert has clarified some of the potential advantages of effect size and suggested some of its applications: both Robert and Ray have pointed out its limitations and have alerted us to some key issues. I would like to thank you both for keeping a non-statistician challenged and informed.

11 Discussion

Michela Gnaldi and Paula Hammond

The use of effect sizes is a keenly debated topic and one that is seen as an important issue in educational research and evaluation for policy making. This chapter discusses the comments made at the invitational seminar, jointly organised by the Institute of Education, University of London and the National Foundation for Educational Research (NFER) on 14 November 2003 and the contributions received from the discussion forum that was available for a month after the event.

11.1 Technical issues

Schagen makes the point that inspite of a general agreement about the effect sizes for binary variables there are still a number of suggestions on how to make equivalent effect sizes for continuous variables. Work by Sammons and Elliot (Chapter 2) and Schagen and Tymms (2003) include a discussion about such calculation methods. The authors suggest a range of multiplying factors on the standard deviation (SD) from 1.0 (Sammons and Elliott, Chapter 2) to 2.0 (Tymms, Chapter 5) with Schagen (Chapter 3) proposing the use of √2 In agreement with Tymms, Strand comments that using a range from 1.0 SD below the mean to 1.0 SD above the mean (i.e. 2.0 SD) seems integral to the standardised concept underlying effect sizes and therefore believes that its use would represent a natural solution to this issue.

Following a further discussion on to how to calculate effect sizes for continuous variables, Sammons now believes that using the 2.0 SD methodology would probably be more accessible than using the cautious 1.0 SD approach previously preferred. She adds that, whichever method is used to calculate effect sizes, it is important that the underlying rationale is made clear in the reporting.

According to Levačić, it is very important to be precise about the definition of 'effect size' because the correct formula to use to obtain the effect size depends on one's definition of the term and on the units in which the variables in the regression equation are measured. She specifies that the term is often used to mean the effect size in standardised units. The effect is then the 'effect' on variable Y measured in standard deviations of a 1 standard deviation change in the variable X. At other times, it refers to the effect of a 1 unit change in variable X on variable Y when both variables are defined in natural units.

On the other hand, Gorard suggests an empirical approach to the calculation of effect sizes and considers the 1.0, 2.0 or 3.0 SD calculation approach to be too rigid. Gorard suggests that if effect sizes are to be valuable, the difference between effect size and measurement error has to be large. Substantial differences rather than a slight variation is what we need to be concerned with. In other terms, differences that outscale measurement errors represent what should be disseminated into policy and practice.

Gorard expresses caution against the use of complex analysis when this is not justified stating: 'if the data can speak don't interrupt'. Schagen supports this opinion and argues that Exploratory Data Analysis (EDA) is an important first stage to any analysis to be able to understand the depth and richness of the dataset. However, he further argues that it is impossible to do any analysis without modelling and that even EDA requires a model. He states that the problem with educational data is that it tends to have a low 'signal to noise' ratio – meaning that there may be important background factors (e.g. prior attainment) which have a large impact and failure to account for these may lead to erroneous models. Schagen states that the conclusions of some simple models may be confirmed or overturned by further complex models, which take into consideration a larger range of background factors and relationships. Schagen provides an example on the impact of class size on pupil attainment to explain his point.

> *Most analysis showed a positive relationship – kids do better in bigger classes... More recent research using sophisticated analysis* (by Blatchford *et al.*) *has shown a positive impact of smaller classes for certain groups – but it took a complex model to show the educationally significant result.*

McNeice and Bidgood agree with Schagen's point and confirm that, whilst EDA is important, there is a place for complex modelling within educational research when such modelling is based on existing coherent educational theory.

Fitz-Gibbon comments that it would be useful to hold a seminar to discuss the widespread use of the $p < 0.05$ level as a criterion for testing statistical significance. She states that this is a particularly serious issue as many people are still surprised that 'counting how many trials have yielded statistically significant findings leads to increasingly wrong conclusions the more data you have'. Sammons agrees with Fitz-Gibbon about the need for greater clarity in discussing effect sizes and their meanings. She argues that proper care and attention is required in reporting research findings in order to stress the nature of the research design used and recognises that over reliance on the $p < 0.05$ may be inappropriate in some instances. McNeice and Bidgood further clarify the point by stating that statistical significance does not tell policy makers or practitioners anything about the size of the effect. In education 'policy makers and practitioners want to know if seemingly significant effects actually have some meaning'.

Fitz-Gibbon suggests that the incautious reader could believe that there were already hundreds of effect sizes in educational research. The reason for this misunderstanding is due to the term effect sizes being used too widely and without differentiation between the sorts of data used to calculate effect sizes. It is therefore not sensible to have the same vocabulary for effect sizes arising from randomised control trials (RCT) and effect sizes arising from 'much less robust designs'. Sammons appreciates that RCTs have a valuable place in educational research. However, she stresses that RCTs raise serious ethical and practical issues, and doubts they should be regarded as a 'gold standard'. She claims that:

> *Policy makers and practitioners have a valid interest in the effectiveness of current provision and variation in quality. Intervention studies can also have limitations because when rolled out different levels of commitment/adherence to design or different contexts may mean different impacts than those identified in an RCT.*

Sammons points out that Fitz-Gibbon's and Tymms' own studies on PIPS, ALIS and YELLIS[1] are not based on an RCT model but provide valuable comparative data. Other educational effectiveness research similarly has provided important evidence of policy relevance without using a quasi-experimental design. However, Fitz-Gibbon responds that whilst epidemiology (PIPS, ALIS and YELLIS) is important, clinical trials are still necessary to ensure that good rather than harm arises from treatments, interventions, policies etc.

[1] For more information regarding PIPS, ALIS and YELLIS see www.cemcentre.org.

11.2 Non-technical issues

Andrew Ray discusses the potential use of effect sizes to inform policy makers and states policymakers are keen to use a wide range of evidence from the United Kingdom and abroad. To understand what works and to spread best practice, there need to be measures of effect size which allow comparisons between different studies based on different policies. On the other hand, Ray also points to the need for a wider agreement on how effect sizes are calculated within different research. He points out that in recent years there has been inconsistent use of the term effect size and the methodologies used to calculate them.

Researchers sometimes quote large apparent effect sizes. Before applying these results in policy formulation we need to ask whether the result is applicable to the current national context and whether achieving such an effect might be prohibitively expensive. We also need to question whether a large effect size is in some way misleading; it could be inflated for example where the intervention is on a relatively homogeneous group of pupils, such as a top maths stream, where the SD of the group will be small. Sammons welcomes Ray's comments on the potential use of effect sizes, his critical questions and awareness that small effect sizes can be considered important. She adds that action on several aspects which show small effects may, in combination, prove more influential than focussing on one aspect which may have a larger effect.

Parker suggests three issues educational researchers need to consider with regard to policy makers.

1. Policy makers are human and are therefore receptive to having their prejudices confirmed and resistant to having them challenged.
2. They all know of historical examples where researchers lied to them or got it wrong, from Cyril's Burt's falsified work on identical twins, through the flawed interpretation of early years research which appeared to show that nursery education hindered educational attainment; to the mischievous use, by Chris Woodhead, of Ofsted data to claim that small class size made no difference to outcomes.
3. When politicians talk about 'evidence based policy making' they mean more than the output of educational research.

Parker states that educational research must be of greatest benefit to policy making when it can challenge erroneous belief, settle debate between differing policy options or offer insight into possible new approaches to old and intractable issues. On the other hand, policy makers require the output from the research community to be both clear and persuasive. He stresses that clear should not be interpreted as 'simple'. The real world complexity in educational research should be retained and demands for simplification should be resisted as 'this could lead to the simplistic'. Parker recognises the benefits of considering effect sizes and wonders whether a combination of better research design and a more consistent approach to effect size considerations would have enabled past research to reach more persuasive conclusions.

Parker points also to the desirability of promoting synthesis and the accumulation of knowledge through the combination and comparison of the results of different studies. He agrees with Goldstein about the necessity for practitioners and policy makers to take responsibility for the necessary contribution of value/significance measures to research findings. Similarly, Melhuish stresses that a key value of effect sizes is the opportunity they provide for comparisons on the size effects for different variables on one outcome, between one variable and different outcomes or, most importantly, between studies.

McNiece and Bidgood suggest that measures of effect size could be calculated to assess the impact of explanatory variables on attainment at successive time points. The use of effect size in longitudinal analysis may help to identify subtle changes in the associations between explanatory variables and educational attainment over time and allow for monitoring the impact of certain factors on progress and development throughout the educational career.

11.3 Summing up

A continuing problem in educational research is that of interpreting the results of statistical analysis in such away that the impact of interventions on educational outcomes can be assessed. Effect sizes has been suggested as the best approach, but its use has been patchy and underlying issues are still debated.

It emerged from the discussion that there is no best effect size measure. Different measures of size effects are required for different studies with different methodologies and different questions. However, it was also highlighted that more agreement on the best approaches to calculate effect sizes under different circumstances is needed.

Earlier in this chapter it was also indicated that relying on the p-value alone when presenting results may be inappropriate and could lead to misreporting. There was a core agreement regarding the need for practitioners and policy makers to combine value/weight/significance measures to research findings. Despite such a general agreement, comments were also made that the role of effect sizes should not be over stressed. The adequacy of models, quality of data and controls made will have a significant impact on the calculated effect size. Good research should address all aspects and publish a range of statistics to enable proper evaluation of any research and interpretation of results.

It was also noted that policy makers and practitioners require the output from the educational research establishments to be both clear and persuasive. Besides, caution is required against over complicated analysis without justification. However, it was also pointed out that simple models may need to be confirmed by more complex models which take into consideration more background factors and relationships.

Overall, it was felt that the seminar will have played an important part in clarifying effect sizes, their use in reporting research findings and a consistent approach in their calculation.

References

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.

BURT, C. (1996). 'The genetic determination of differences in intelligence: a study of monozygotic twins reared together and apart', *British Journal of Psychology*, **57**, 1, 137–53.

Acknowledgements

The authors would like to thank the participants in the seminar discussion.

Penny Bidgood, Kingston University
Karen Elliot, Institute of Education University of London
Carol Fitz-Gibbon, University of Durham
Harvey Goldstein, Institute of Education University of London
Stephen Gorard, University of Cardiff
Ros Levačić, Institute of Education University of London
Rosemary McNeice, Kingston University
Ted Melhuish, Birkbeck, University of London
Alan Parker, NFER Board of Trustees
Andrew Ray, DfES
Pam Sammons, Institute of Education University of London
Ian Schagen, NFER
Steve Strand, nferNelson
Peter Tymms, University of Durham

12 Final summary

Ian Schagen and Karen Elliot

'But what does it *mean*?' was chosen as the title of this volume, not because we expected to answer this question, but because it encapsulates the concerns often expressed by the users of educational research that they are unable to see the practical and policy wood for the methodological trees. This is a long-term issue, but we believe these proceedings have facilitated the ongoing process of debate, research and theoretical and practical advancement in this field of enquiry. As a starting point for the next stage of the process, it is useful to summarise the key issues that arose and provide indicators of potential areas for further developments.

In some ways, the concept of 'effect size' has been appropriated by educational research from other fields where it is more commonly understood, in a similar way to that in which 'value added' was lifted from economics some years earlier. The idea of an effect size is more familiar to those working in areas where controlled experiments are employed, such as clinical trials – with due respect to Fitz-Gibbon (2004), social and educational research does not commonly employ such methods, nor are they likely to replace surveys and the analysis of administrative datasets in the near future. So the challenge for educational researchers and others is to interpret this concept in ways which are relevant to our field. As discussed in earlier chapters, the terminology itself is slightly unfortunate, as 'effect size' has overtones of causality which are more appropriate to experimental findings than to those obtained from surveys or modelling naturally occurring variation. It might have been better to have coined a new phrase to express such a statistic within this particular context – something like 'standardised coefficients' – but this is probably not now an option.

So in what situations might this concept be used within educational research? A number of potential applications have been identified in this publication, including:

1 To enable coefficients of different variables within complex models of educational outcomes to be directly compared with each other in terms of 'strength' or 'importance'.

2 To enable such measures to be compared across different studies in ways which are independent of the exact units used in each.

3 To enable findings from complex analyses to be presented to a non-technical audience (such as policy-makers) in ways which are accessible and speak directly to their agendas.

The first two purposes are particularly important but give rise to a fairly substantial number of technical questions, some of which have been addressed in this volume.

- What is the correct estimate of standard deviation to use in computing effect sizes?
- Should there be a factor of 2, or 1, or something in between, to multiply the standard deviation?
- How should effect sizes for binary and non-binary variables be compared?
- How are confidence intervals for effect sizes calculated?
- How can the concept be extended to other situations, for example as discussed in earlier chapters variance ratios (Strand) and interaction terms (Schagen)?

But in all this, we must not forget Godfrey's warning that effect size is a 'pseudo-concept', not a direct measure of an underlying reality but a way of expressing the results of our (inevitably partial and incomplete) models of that reality. Complex and technical discussions are an important part of the professional remit of researchers and statisticians, but they are not the only part of that task. If the results from analyses are not conveyed in ways that speak clearly to the intended audience, then we have failed.

For this reason the search for clear but valid ways of presenting results must assume one of the highest priorities. Examples have been produced in the papers in this volume of how this might be done, but there is obviously a great deal yet to do in terms of improving, standardising and disseminating best practice in presentation. As part of this, it is possible that the pendulum may swing away from the dimensionless purity of effect sizes towards measures which are expressed in more natural units. The search for units which carry broad currency across the educational field needs to continue, together with the development of robust ways of converting modelling results into such units.

In summary, we hope that the seminar and this publication are the beginning of a process in educational research which will bring together a range of different practices in the search for some common standards in the presentation of complex modelling results to answer the question 'But what does it mean?'

Reference

FITZ-GIBBON, C. (2004). 'Editorial: the need for randomized trials in social research', *Journal of the Royal Statistical Society A*, **167**, 1, 1–4.

Glossary

Age-standardised score

Test scores adjusted to give values which are evenly distributed about a fixed value (often 100), and which take account of the age of the individuals taking the test.

Binary variables

A variable that can take only two values, representing, for example, sex, absentees etc. Also known as Dichotomous variables.

Causality

The relationship between cause and effect. The principle that all events have sufficient causes.

Coefficient

In *regression analysis*, the estimated relationship between the outcome measure and one of the background variables, expressed as the change in the value of the outcome associated with one unit change in the background variable.

Confidence interval

No statistical estimate is ever totally accurate, but the degree of accuracy will depend on a number of factors, including the amount of data on which it is based. If we estimate something, say the overall national mean score on a test based on a sample of individuals, we may compute a confidence interval that is a range of values, enclosing the best estimate, which has a specified probability of containing the true value. For example, if we say that a 95% confidence interval for the mean score is 15.7 to 16.1, this implies that if we repeated the whole exercise lots of times with different samples, then 95 times out of 100 the true national mean score would lie inside the confidence interval we give. Obviously, the higher the confidence level we want, the wider the interval we must specify.

Continuous variable

A variable which is considered to be measurable on a continuous scale e.g. height, weight, test scores etc.

Correlations

A measure of association between two measurements, e.g. between size of school and the mean number of GCSE passes at grades A, B & C obtained by each pupil. A positive correlation would occur if the number of passes increased with the size of the school. If the number of passes decreased with size of school there would be a negative correlation. Correlations range from -1 to +1 (perfect negative to perfect positive correlations); a value of zero indicates no linear association between the two measures.

Dichotomous variables

A variable that can take only two values, representing, for example, sex, absentees etc. Also know as Binary variables.

Effect size

A statistic, often abbreviated to D or delta (Δ), indicating the difference in outcome for the average subject who received a treatment from the average subject who did not. This statistic is often used in meta-analysis.

Explanatory variables

Variables which can be used to explain an outcome. Also known as background or independent variables.

Likert Scale

This scale measures the extent to which a person agrees or disagrees with a question.

Median

The central value in a set of data, such that half the cases lie below and half above that value. It is less affected by extreme values than the *mean* as a measure of the 'average' of a dataset.

Meta-analysis

Meta-analysis is the combination of data from several studies to produce a single estimate. From the statistical point of view, meta-analysis is a straightforward application of multifactorial methods.

Multilevel modelling/analysis

Multilevel modelling is a recent development of linear *regression* which takes account of data that is grouped into similar clusters at different levels. For example, individual pupils are grouped into year groups or cohorts, and those cohorts are grouped within schools. There may be more in common between pupils within the same cohort than with other cohorts, and there may be elements of similarity between different cohorts in the same school. Multilevel modelling allows us to take account of this hierarchical structure of the data and produce more accurate predictions, as well as estimates of the differences between pupils, between cohorts, and between schools. (Multilevel modelling is also known as hierarchical linear modelling).

Multivariate analysis

The analysis of data, which is multivariate in the sense that each member bears the values of p variates.

Noise

A series of random disturbances. Noise results in the possibility of a signal sent, x, being different from the signal received, y.

Normalised scores

In the analysis of data it is often desirable to convert each set of original scores to some standard scale. This process is known as the normalisation of scores.

Null hypothesis

This term relates to a simple hypothesis of no change or difference, as distinct from the alternative hypothesis of a significant difference that is being tested.

Percentiles

The set of partition values which divide the total frequency into one hundred equal parts.

p-value

The probability that, given the null hypothesis, a particular statistic takes a value at least as extreme as that observed. Often for a statistical test we require that a p-value be smaller than some fixed value, if we are to reject the null hypothesis.

Quartiles

Numerical values which divide a dataset into four parts with equal numbers in each. The first quartile is such that 25% of the data lies below it and 75% above; the third quartile has 75% below and 25% above. The second quartile is equivalent to the median.

Regression analysis (linear)

This is a technique for finding a straight-line relationship which allows us to predict the values of some measure of interest ('dependent variable') given the values of one or more related measures. For example, we may wish to predict schools' GCSE performance given some background factors, such as free school meals and school size (these are sometimes called 'independent variables'). When there are several background factors used, the technique is called multiple linear regression. If just a single background factor is used to predict, we have simple linear regression, and the results may be plotted as a straight line on a graph.

Reliability

The 'reliability' of an outcome is a measure of the extent to which it is due to permanent systematic effects, and therefore persists from sample to sample.

Standard deviation

Standard deviation is a measure of the spread of some quantity within a group of individuals. If the quantity is distributed approximately Normally, we would expect about 95% of the individuals to be within 2 standard deviations either side of the mean value.

Statistical inference

The extension of sample results to a larger population.

Statistical significance

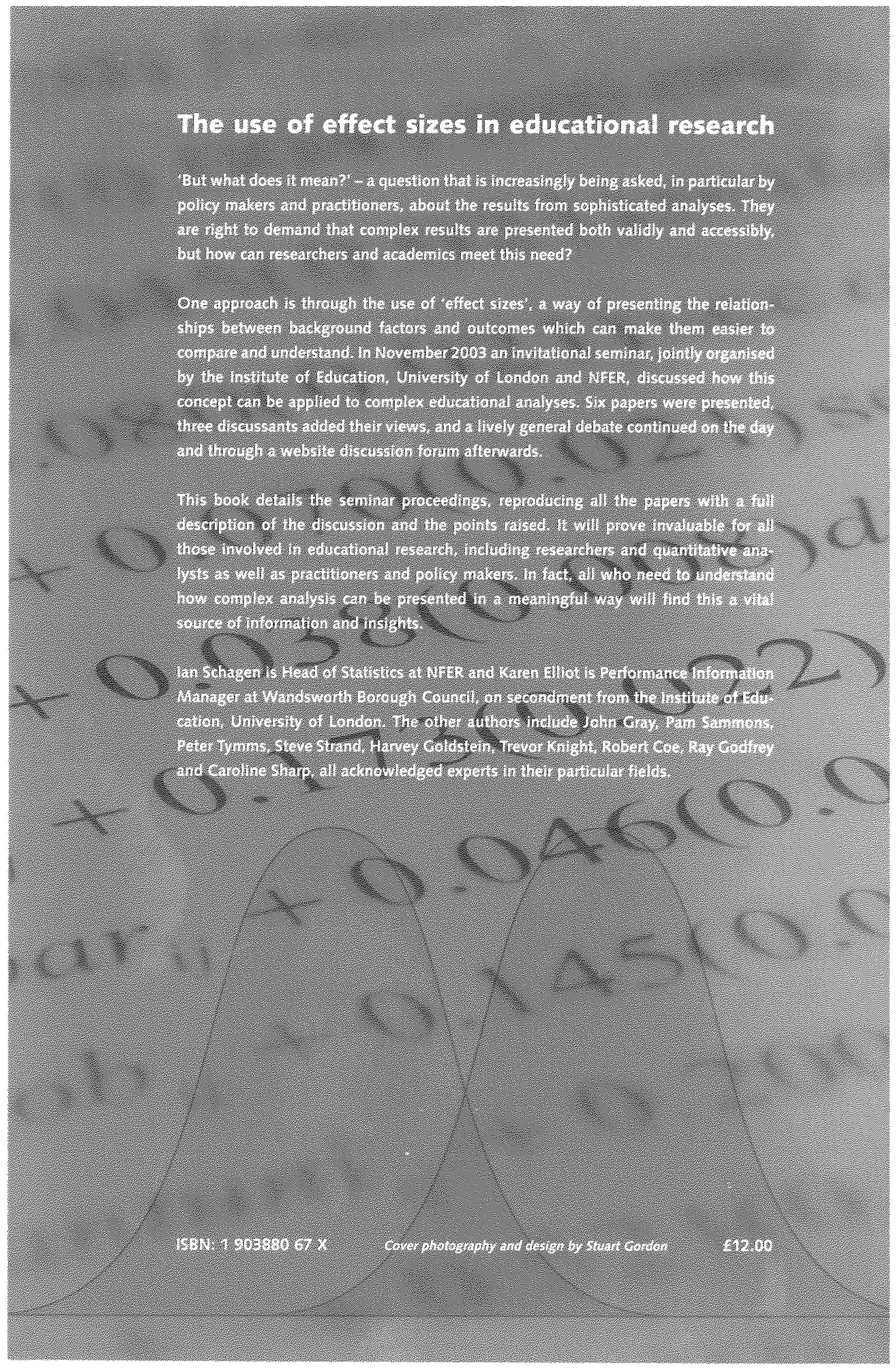
We say that there is a statistically significant difference between two groups in some quantity if the probability of that difference arising by chance is less than a preset value (e.g. 5%). Similarly, we say that there is a significant relationship between two variables if the observed results have a low probability of arising by chance, which is by random fluctuations when the two variables are really unrelated.

Variability

The quality of being likely to change or vary over time; lack of uniformity.

Variance

A measure of variability in data (the square of the *standard deviation*).



The use of effect sizes in educational research

'But what does it mean?' – a question that is increasingly being asked, in particular by policy makers and practitioners, about the results from sophisticated analyses. They are right to demand that complex results are presented both validly and accessibly, but how can researchers and academics meet this need?

One approach is through the use of 'effect sizes', a way of presenting the relationships between background factors and outcomes which can make them easier to compare and understand. In November 2003 an invitational seminar, jointly organised by the Institute of Education, University of London and NFER, discussed how this concept can be applied to complex educational analyses. Six papers were presented, three discussants added their views, and a lively general debate continued on the day and through a website discussion forum afterwards.

This book details the seminar proceedings, reproducing all the papers with a full description of the discussion and the points raised. It will prove invaluable for all those involved in educational research, including researchers and quantitative analysts as well as practitioners and policy makers. In fact, all who need to understand how complex analysis can be presented in a meaningful way will find this a vital source of information and insights.

Ian Schagen is Head of Statistics at NFER and Karen Elliot is Performance Information Manager at Wandsworth Borough Council, on secondment from the Institute of Education, University of London. The other authors include John Gray, Pam Sammons, Peter Tymms, Steve Strand, Harvey Goldstein, Trevor Knight, Robert Coe, Ray Godfrey and Caroline Sharp, all acknowledged experts in their particular fields.

ISBN: 1 903880 67 X

Cover photography and design by Stuart Gordon

£12.00