*Effective Education*
Vol. 3, No. 2, September 2012, 1–19

Routledge
Taylor & Francis Group

# Educational effectiveness: the development of the discipline, the critiques, the defence, and the present debate†

David Reynolds[a], Christopher Chapman[b]*, Anthony Kelly[a], Daniel Muijs[a] and Pam Sammons[c]

[a]*School of Education, University of Southampton, Southampton, UK;* [b]*School of Education, University of Manchester, Manchester, UK;* [c]*Department of Education, University of Oxford, Oxford, UK*

(*Received 22 November 2011; accepted 14 April 2012*)

Educational effectiveness research (EER) has made a significant contribution to our understanding of the characteristics and processes associated with more and less effective schools in a diverse range of contexts. However, this remains a contested field of inquiry and has been subjected to significant critique. This paper examines the origins and development of EER and summarises the key critiques and defences of the field during the past 30 years. It then moves on to examine the recent critique of the field by Stephen Gorard in the UK and responds by highlighting statistical errors and simplistic claims made by Gorard about the field's involvement with the development of national value-added systems and interaction with policy-making in his recent papers.

**Keywords:** educational effectiveness; school improvement; multi level modelling



## Introduction

Educational effectiveness research (EER) is concerned with understanding the key factors and their interactions that lead to more or less effective classrooms, schools and education systems. As a field of inquiry, it is relatively young, but has developed on a steep theoretical and methodological trajectory. The origins of the field can be traced back to responses to sociological research and policy that denied that schools could make a difference to educational outcomes and ultimately the life chances of young people. Over the past 30 years, however, EER in many countries has consistently demonstrated that teachers and schools can and do make a difference. Despite these findings, there have been a number of critiques and counter critiques relating to the field, including a recent series of criticisms by Stephen Gorard. These are worthy of particular attention because they are based on statistical errors and a simplification of the relationship between EER and education policy-making.

The paper begins by providing an overview of the development of EER before then reflecting on the key critiques and defences of the field. Following this, the paper turns

*Corresponding author. Email: chris.chapman@manchester.ac.uk

†This is a substantially enlarged and revised version of a paper originally entitled 'Stephen Gorard and School Effectiveness: A Rejoinder and Final Response' presented to the Annual Meeting of the British Educational Research Association, London, September 2011.

ISSN 1941-5532 print/ISSN 1941-5540 online

http://dx.doi.org/10.1080/19415532.2011.686168
http://www.tandfonline.com

to the current set of critiques made by Gorard, taking each in turn and examining their validity in detail. In conclusion, the final section provides a succinct summary of the key points provided in this paper.

## The development of EER

The development of EER as a field of inquiry can be traced back to the 1970s. It emerged as the first serious critique to sociologically driven beliefs of the 1960s in which the structures of society were considered to be so strong in their influences that the socio-economic circumstances a child was born into determined their life chances and general futures. It was thought education and particularly schools could have little influence on social mobility and that 'schools made no difference' (Bernstein, 1968).

The analysis within *Equality of educational opportunity* published by Coleman and his colleagues typified the mood of the time. The report concluded

> Schools bring little influence to bear on a child's achievement that is independent of his [sic] background and general social context; and that this very lack of an independent effect means that the inequalities imposed on children by their home, neighbourhood and peer environment are carried along to become the inequalities with which they confront adult life at the end of school. (Coleman et al., 1966, p. 325)

Regression analysis in this study suggested that the prime determinant of a pupil's academic success was his or her socio-economic background, and therefore the common assumption of the time became that schools could do very little to impact on student outcomes, as academic achievement appeared to be predetermined by factors external to the control of schools. This report provided pessimistic reading for those involved in education, probably affecting the morale of teachers. However, while this report attracted attention for its findings, there was also criticism of the methodology used (see Teddlie & Reynolds, 2000). Nevertheless, further studies in the USA, for example, that by Jencks et al. (1972) which reanalysed the Coleman et al. (1966) data, also supported the initial findings, thus adding credibility to the assertion that schools could make little difference to students' learning and therefore did not really 'matter'.

At the same time, the Department of Education and Science (1967) in the UK, published the Plowden Report. This report also highlighted the limited contribution of the school compared to that of parental attitudes and home conditions on student outcomes. Other international studies of this era also supported the conclusions from the Plowden Report, and it was because of this, and the American research by Coleman and others proceeding at the same time, that an international acceptance emerged that schools could only have a marginal influence on academic outcomes.

At the policy level, it was not until 1976 when the then Labour Prime Minister in the UK, James Callaghan, addressed a meeting at Ruskin College, Oxford, on the issue of education that the potential for schools to make a difference was given credible attention. Callaghan outlined the challenges for British education, demanding a rise in standards and greater accountability from schools. He, implicitly if not explicitly, saw that educational systems made a difference. Three years later, Rutter, Maughan, Mortimore, and Ouston (1979) published *Fifteen thousand hours* – the book so titled because from

the age of five this was the time that a child spent in compulsory schooling. Rutter et al. (1979) concluded that, after accounting for intake differences,

> Students at the most successful secondary school got four times as many (examination) passes on average than those at the least successful school. (p. 19)

The study also identified a number of common factors associated with more and less successful schools. Schools found to be generally effective showed benefits across the whole range of student academic and social outcomes, rather than only in a few areas, suggesting a consistency of school effects. Despite using a cohort design that matched individual pupil data at intake to school and at outcome at the age of 16, the study was harshly criticised (Goldstein, 1980), which reflected the prevailing socio-political climate that remained pessimistic in terms of the extent that schools could make a difference to student outcomes. A major criticism levied by those such as Cuttance (1982) covered the attempt to generalise the findings to all schools with his criticisms focusing on the size and nature of the particular sample; the study involved only 12 schools and only 2000 pupils. The London inner-city schools picked were also considered not representative of the wider population of secondary schools to be found in England.

Despite the criticisms, for the first time in the UK, a major study had indicated that schools could make a difference to student learning. This first phase work presented a significant challenge to the sociological beliefs and explanations that 'schools didn't matter'. Following the work of Rutter and colleagues, other small-scale studies reported similar conclusions about the size of school effects (e.g. Reynolds, 1976; Reynolds & Sullivan, 1981). These studies involved collecting a large range of data on areas such as children's attitudes to school, organisational school factors, and cultural factors within schools. The results from these studies suggested that there were a number of factors associated with more effective schools, including positive academic expectations and high levels of student involvement (e.g. Reynolds, 1976). This second phase of school effectiveness research attracted harsh criticism, often related to its methodology which used not matched pupil cohort studies, but group-based, cross-sectional data on intakes and outcomes.

In the third phase, researchers worked to develop more sophisticated methodologies, including the development of multilevel statistical analysis. This phase was led by Aitkin and Longford's (1986) re-analysis of Gray's (1981) data that used multilevel modelling (MLM) for the first time in the estimation of school effects. Two of the most important studies in the UK were also undertaken in this phase by Mortimore, Sammons, Stoll, Lewis, and Ecob (1988) and Smith and Tomlinson (1989). Mortimore and his colleagues used 50 randomly selected London primary schools from a total of over 650, a sample of schools that was later found to be the representative of schools throughout London. In their findings, they reported on a number of their schools that were effective academically and socially. These schools possessed the following characteristics:

- purposeful leadership of the staff by the head teacher,
- the involvement of the deputy head,
- the involvement of teachers,
- consistency among teachers,
- structured sessions,

- intellectually challenging teaching,
- a work-centred environment,
- a limited focus within sessions,
- maximum communication between teachers and students,
- record keeping,
- parental involvement,
- a positive climate.

The study also concluded that the school was 4 times more important than pupil background factors in accounting for pupil progress in reading and 10 times more important in mathematics (Mortimore et al., 1988, pp. 186–188).

Smith and Tomlinson (1989) reported variations in effectiveness between 18 comprehensive secondary schools in the UK. For some groups of students, the variation between individuals in different schools accounted for a quarter of their total variation in examination results. They also reported some 'differential effectiveness': out of 18 schools, the school ranked most effective for mathematics was ranked only 15th most effective for English attainment, after allowance for intake differences. Teddlie and Reynolds (2000) summarised the general findings from this study as:

- The overall percent of variance in achievement explicable at the school level was around 10 per cent across all subjects, ability groups and ethnic groups;
- The effect of the school varied by the achievement level of students, with the effect being less for average achieving students;
- There was a variation in the effects of different Departments;
- There was a small differential school effect for students from different ethnic groups, with some schools doing better than others. (p. 87)

In England, the 1992 Education (Schools) Act gave the government the power to nationally publish performance tables of schools' academic performance, leading to elements of the media ranking secondary schools in crude league tables based upon their examination results. Standardised national inspection by the Office for Standards in Education (OfSTED) also began in the 1990s and the criteria used by OfSTED to judge the effectiveness and quality of schooling and effective school processes can be traced back directly to the early studies that characterised elements of effective schooling, as in the Sammons, Hillman and Mortimore (1995) review of the field.

The use of crude examination results to rank schools, combined with an inspection system underpinned by a focus upon 'key characteristics', provided the catalyst for some researchers within the educational effectiveness community to develop more sophisticated mechanisms to examine school effectiveness. Fitz-Gibbon (1992, 1996, 1997) and her colleagues invested much time and energy in this area and developed data systems based on the concept of the self-monitoring educational system. In an attempt to move beyond the 'whole school' level, other researchers were keen to explore different levels of analysis and Harris, Jamieson, and Russ (1995) and Sammons, Thomas, and Mortimore (1997) characterised common traits of more effective departments and Creemers (1994) began to explore classroom effects.

Interest in the effects of the classroom and of teachers continued into the 2000s with Muijs and Reynolds (2000) and Day, Sammons, Stobart, Kingston, and Gu (2007) further examining teacher effects. The 2000s also saw ever more sophisticated

approaches to assessing student progress developed and a number of researchers including Van Damme, Van Landeghem, Onghena, De Fraine, and Opdenakker (2001) and Van de Gaer, De Fraine, Van Damme, De Munter, and Onghena (2009) became concerned with assessing a broader range of outcomes. Issues of equity also came to the fore (see Sammons, 2010), which involved some researchers turning their attention to the differential effects of schools upon different groups such as the white working class and African Caribbean boys and differences in performances between gender groups and children of different socio-economic status (Strand, 2010). Others explored within-school variation and the differential effects within schools at departmental and teacher levels (Reynolds, 2007).

Most recently, methodologies have become even more sophisticated and we have seen a growing interest in the use of randomised controlled trials in the UK (e.g. Connolly, Miller, & Eakin 2010; Miller, Connolly, & Maguire, 2011). MLM techniques have also been used to compare the impact of innovative teacher training programmes such as *Teach First* (Muijs et al., 2010; Muijs, Kelly, Sammons, Reynolds, & Chapman, 2011) and government interventions including *The Extra Mile* (Chapman, Mongon, et al., 2011) and the effectiveness of various structural arrangements such as federations (Chapman, Muijs, & MacAllister, 2011). The most recent thinking in the field are the findings from an Economic and Social Research Council-sponsored seminar series which focused on challenging the orthodoxy of school effectiveness research, which sets out a new agenda (see Chapman et al., 2012), recommending:

- an enhanced focus upon context specificity and the effectiveness factors operating in different contexts;
- understanding the level of the school and the interactions between the school and classroom levels;
- development of a broader range of outcome measures;
- utilising insights from cognitive neuroscience;
- enhanced links with other disciplines;
- strengthening of the theoretical basis of the field;
- continuing the methodological advances.

## The critiques and the defences

We have so far provided a context for this paper by offering a brief overview of the origins and development of our field. A more comprehensive analysis can be found in the two standard handbooks on this subject: Teddlie and Reynolds (2000) and Townsend (2007). We now turn our attention to the key critiques and defences associated with EER over the years of its development.

A wide variety of critiques of EER have emanated over the years, no doubt in response to its emergence as a discipline from nowhere as it were and to the evidence that it was being heavily used by policy-makers (see Reynolds, 2010, for a review), and also to its popularity as a novel explanation for educational failure. These criticisms have been analysed elsewhere (Teddlie, 2009) and indeed the early criticisms formed the basis for a well-known debate at the American Educational Research Association in 2000 between the critics (Slee, Thrupp, and Weiner) and the discipline's advocates (Teddlie and Reynolds). The debate generated a special issue of the journal *School Effectiveness and School Improvement* (Reynolds & Teddlie, 2001; Slee & Weiner, 2001; Teddlie & Reynolds, 2001; Thrupp, 2001; Townsend, 2001), itself an interesting

reflection on how EER researchers have tried to consider criticisms rationally and learn from them, a characteristic not always evident in our critics.

If we look at the criticisms in detail, and assess their validity, we can identify four key themes. First, early work was seen to rely too much upon qualitative case studies of 250 'effective schools' that did not possess the methodological rigour to definitely detach school effects from pupil intake effects (Cuban, 1983; Cuttance, 1982; Good & Brophy, 1986; Purkey & Smith, 1983). These criticisms have been answered by the development of much more rigorous designs for EER that collect multiple data to 'control out' non-school factors (see Creemers, Kyriakides, & Sammons, 2010, for a 255 sample of studies).

Second, EER was argued to over-emphasise the influence of schooling rather than the more determinate influences of the social class (Slee, Weiner, & Tomlinson, 1998; Thrupp, 1999; Willmott, 1999; Wrigley, 2004), yet the most recent research in EER which employs the most sophisticated methodological and statistical methods shows 260 much higher 'effects of schools' than the early 12–15% of variance explained that the critics highlighted. Guldemond and Bosker (2009), for example, show school-level variance explained 30% to 50% of variations between students, and Luyten, Tymms, and Jones (2009) over 33%, both figures considerably in excess of earlier estimates and both similar to family background effects.

265 Third, EER was widely criticised for neglecting to generate theoretical analyses that could link the empirical findings together (Elliott, 1996; Goldstein, 2008; Slee et al., 1998; Wrigley, 2004), a completely valid criticism. Recent research in EER has been notably successful in this area with the development of the Creemers and Kyriakides (2008) dynamic theory of educational effectiveness which is now being tested in mul-270 tiple studies in multiple sites.

Finally, EER was widely criticised (Slee et al., 1998; Thrupp, 2001) for generating 'one-size-fits-all' models that did not vary with context, in terms of the proposed educational factors responsible for pupil outcomes. However, although a sensible criticism, it needs to be acknowledged that this was a consequence of the over-sampling of dis-275 advantaged communities because of the desire of EER researchers to understand and help disadvantaged children, a political and social orientation very far from that alleged by many of the critics. In any case, as time has gone on, there has been increasing awareness of this issue and burgeoning research interest in EER (e.g. Chapman et al., 2012; Teddlie & Reynolds, 2000).

280 It seems, then, that EER has developed in ways that would answer many of the criticisms levelled at us by our critics. This is not surprising since there is much evidence of continued self-criticism, self-evaluation, and reflexivity among EER researchers (see e.g. Chapman et al., 2012; Reynolds, 2010; Teddlie & Reynolds, 2000; Townsend, 2007). The Townsend (2007) handbook had a chapter explicitly based on the criticisms 285 of the field and their possible validity by well-known critics from outside the field. The most recent *International handbook of educational effectiveness and improvement research* (Chapman, Muijs, Reynolds, Sammons, & Teddlie, 2013) has one chapter on critiques of the field, and reflection on the criticisms is in evidence in virtually all the others.

290 Interestingly, probably as a recognition of the attempts, we have made in EER to listen and respond to critiques with more valid and appropriate research designs, the flurry of criticisms that emerged in the 1990s and early 2000s had become muted in recent years and much less prevalent. But, then, in 2010, Stephen Gorard began his particular examination of EER and finds us wanting. The recent criticisms are

worthy of a response on two counts. First, Gorard's claims are based on statistical errors and, second, the arguments presented conflate the field of inquiry with government policy. It is on this basis we assert this critique as invalid and overly simplistic.

## The present debates

The recent debates were triggered by an article in the *British Educational Research Journal* (*BERJ*) by Gorard (2010a) to which we responded in *Research Intelligence* (*RI*) (Muijs et al., 2011). The same issue also carried a rejoinder by the author of the initial article (Gorard, 2011a) and a subsequent further paper entitled 'Comments on the value of educational research' which summarised the critique of our initial response while widening the attack to other issues related to EER. This paper was uploaded onto the conference website for the British Educational Research Association Conference of 2011 (Gorard, 2011b).

This section of our paper is structured in three sections. The first relates to methodological and statistical issues in the *BERJ* paper (Gorard, 2010a), the cornerstone of the current debate, and the Gorard (2011b) paper. The second relates to Gorard's broader attacks on EER, and third, we reflect on the highly personal nature of the attack upon us and the field, the potential damage to the reputation of educational research in general that such behaviour may cause, and the need for educational researchers to avoid such behaviours. We then conclude the paper with a summary of our argument.

### *Statistical and methodological issues*

#### *The issue of relative error*

Gorard's (2010a) basic thesis is that the impact of measurement errors in any data used in school effectiveness studies means that both the variance attributable to the school level (measured by the intra-school correlation) and also particularly the customary individual school residual scores that have been used to provide indicators of differences between individual schools in their 'effectiveness' in promoting pupil progress are so small and unreliable that they should not be used as measures of school effects (as is the case with the historical use of published school performance tables for example).

There are a number of points worth making in response to this argument. First, the papers themselves are confused about these statistical matters. They use various terms – 'error range', 'relative error', 'relative error range', 'maximum relative error', and 'propagated error' – interchangeably in the three papers, though each of these have clear and distinct meanings. The papers also contain some serious errors in the calculation of relative error, which are concealed by changing the terminology used in the later Gorard (2011a) article. In this article, it is argued that we do not understand what a maximum relative error is and therefore 'miss the enormity of the problem' (line 5). Specifically:

> The relative error is in proportion to the number in which the error occurs. So the full range for the calculated CVA[1] of 1 in their example is 20 (from −9 to +11). This is 20 times, or 2000% of the CVA score of 1. (Line 18)

But the measurement error of 20 is of course related to the contextual value-added (CVA) score of 100, so the percentage relative error is in fact 20% not 2000%. The *range* for the calculated CVA is certainly 20 times the CVA score of 1, which would be 2000% if that had any meaning, but this is not *relative error*.[2] If a pupil were to score 100 with instead an estimate of 99.99 say, with the same maximum error range of $\pm 10\%$ as Gorard wishes, is he seriously suggesting that although nothing has changed with the measurement or with the methodology or with the error range, that the relative error in school effectiveness research has now grown to 200,000% (i.e. 20/0.01). This is a *reductio ad absurdum* and would mean that if a pupil scored 100 on both measurement and estimate, there would be an infinite relative error [20/0] in the same measurement as produced 2000 a week ago and 200,000 a minute ago. Relative error in measurement cannot behave like this for it to have any meaning. This is why relative error is derived by dividing the error by 100 and not 1, or 0.01, or zero. Furthermore, 20% is a 'maximum' error, which is important because since errors are likely to be normally distributed, actual errors will in most cases be very much smaller. In fact, the original paper (Gorard, 2010a) claimed that '. . .the maximum relative error . . . is a massive 3,980%' because he had wrongly confused a predicted score with a measured score, but in the *RI* article this is subtly altered to read 'the relative error range is 3,980%' (line 19).

To summarise, there are some major statistical errors in calculating the potential impact of measurement error in his analysis. Moreover, in discussing the possible impact of measurement error on student prior attainment scores (and a similar issue would apply to student outcome scores of course since all measurements have a margin of error), the two papers ignore the fact that such errors tend to be randomly distributed. Thus in calculating individual school residuals based on results for a group of students nested at the school level, the over- or under-measurement of individual student's attainments at different time points tend to cancel each other out and is unlikely to be systematically different in different schools, so therefore the relative size of residual school estimates would be unchanged. Moreover, many of us within EER have consistently advocated the use of confidence intervals (CIs) associated with individual school residuals, and cautioned against any attempts to rank schools' value added or indeed their raw performance (both in journal articles and reports to government), yet Gorard ignores these careful caveats often made about how we should interpret differences in school performance (e.g. Goldstein & Spiegelhalter, 1996; Mortimore, Sammons, & Thomas, 1994; Nuttall, Goldstein, Prosser, & Rasbash, 1989; Sammons, Hillman, & Mortimore, 1995; Sammons, Nuttall, & Cuttance, 1993). Interestingly, Goldstein (2008) provides an historical policy analysis and discussion of the way CIs were eventually included in the Department for Education and Skills (DfES) CVA measures in England but notes how the importance of statistical uncertainty as measured by CIs has been largely ignored by politicians and the media, who continue to lay more emphasis on simple raw benchmarks and league tables of schools' raw results. Q5

It is wrong to suggest that EER has ignored the potential impact of measurement error. Indeed, several studies have examined this but have drawn different conclusions from Gorard (e.g. Ferrao & Goldstein, 2008). Adjusting for the measurement error in measures of student prior attainment (as Gorard advocates) tends to *increase* the size of the overall school effect (measured by the intra-school correlation) rather than reduce it and the whole question of measurement error has received attention from researchers studying school effects. Fifteen years ago, Woodhouse, Yang, Goldstein,

and Rasbash (1996) showed that adjusting for the measurement error of prior attainment scores at the pupil level made more difference to calculations of the relative importance of other student-level predictors, such as free school meals (FSM) eligibility, but little difference to estimates of school effects at level 2. Rather, such adjustment reveals an increase in the relative size of the overall school effect measured by the intra-school correlation:

> The level 1 residual variance decreases markedly when adjustment is made for measurement error in an explanatory variable at level 1. This produces a substantial effect on the intra-school correlation, which is further increased when adjustment is made for measurement error in the response variable. (Woodhouse et al., 1996, p. 211)

This finding has since been confirmed in a range of studies (e.g. Ferrao & Goldstein, 2008; Guldemond & Bosker, 2009). Moreover, Goldstein, Kounali, and Robinson (2008) examined the impact of adjusting for measurement error in prior test scores and other binary predictors like FSM eligibility and argued

> Substantively, we can conclude that moderate amounts of measurement error and small misclassification probabilities only result in small changes to parameter estimates. With large errors the effects are noticeable, but are confined in the fixed part of the model to those predictors with error. The level 1 variance estimate, however, is sensitive to the reliability assumed. In particular, the coefficient estimate for free school meals is changed noticeably for a small measurement error variance and the given misclassification probabilities. (p. 256)

Interestingly, in his own work on school composition, Gorard ignores the potential impact of measurement error on predictors such as the percentage of pupils in a school eligible for FSM which may well be more prone to measurement error than measures of attainment. Gorard has gone as far as to claim that FSM is a highly reliable indicator when he uses the National Pupil Database (NPD) for his own research on schools, even though in his 2010a article he raises questions about the value and reliability of many measures in the NPD.

### *The issue of random sampling*

Gorard (2011b) asserts that 'anyone conducting significance tests with a non-random sample is flouting a fundamental assumption of statistical analysis' (p. 11, line 20). It is a difficulty that, if accepted literally and strictly, casts doubt on most, if not all, quantitative social science research, both national and international, not just educational research and not just EER. Achieving truly random samples is rarely possible, which is why most researchers typically do not often use them, but the incremental growth in our understanding of educational and social life generated in fields like EER by using non-random, though usually broadly representative, samples over the last decades has been huge. Moreover, results from EER studies that have used random or stratified random samples of schools have produced broadly consistent findings to those using other samples (e.g. all those schools in a local authority (LA) or a national database). It is simply naive to infer from this difficulty that the great majority of existing educational research is of no use. This view of statistical method is an example of a limited and rather old-fashioned view of statistics, which sees inference as referring only to finite populations, rather than the more popular, contemporary model-based views concerning inference,

which takes the data at hand and attempts to formulate and evaluate the structure of relationships. This is an old argument, which has already been extensively refuted elsewhere (e.g. Plewis & Fielding, 2003), but which Gorard apparently insists on revisiting. For example, in a direct response to similar arguments about the supposed inappropriateness of using statistical significance tests on a population data set for schools and pupils, Goldstein and Noden (2004) pointed out that

> Gorard's third point exemplifies a common misunderstanding of the nature of statistical modelling as applied to social data. He argues that since we have the 'total population' of the schools being studied for the period of concern, probability statements, including significance tests, are irrelevant. On the contrary, social scientists are really interested in the underlying processes that produce the observed outcomes. Thus, for example, suppose we allocated pupils to schools in a purely random fashion and calculated the proportions eligible for free school meals in each school. We would certainly observe differences, but these would have arisen purely as a result of random sampling even though we would have counted the whole 'population'. Thus, in making comparisons between schools we must take account of such sampling variation and this is precisely what statistical models of social processes do. (p 441)

*The issue of MLM*

EER has relied upon MLM as its methodology of choice in much of the recent quantitative work in the field, although of course many other techniques are useful where appropriate (for a discussion of methodological advances in EER, see Chapman et al., 2012; Creemers et al., 2010). Gorard has previously argued that MLM offers few advantages for educational research (2007) and he continues this train of thought in his recent paper, claiming MLM is an 'alleged whizzo cure' (p. 22, line 31) and 'the claimed benefits of multilevel modelling apply, if they apply anywhere, only to random sampling' (p. 22, line 31).

It is unclear to us why Gorard wishes to revisit this well-trodden ground, rehearsing this view yet again, since the arguments in defence of MLM were convincingly marshalled by Goldstein (2003), Plewis and Fielding (2003) and others only a few years ago. We echo Goldstein and Noden's (2004) conclusions about Gorard's critique of MLM, which they argue demonstrates a lack of understanding of the statistics underpinning its use and the increasingly wide application of MLM in the social sciences to analyse and address the complexity of social reality. In EER, this complexity includes the study of different sources of influences on children and students' outcomes, taking a longitudinal perspective that focuses on studying change over time and reflecting the clustering inherent in educational data sets.

The key reason for using MLM is that it provides us with a more accurate reflection of the nested or clustered structure of relationships in educational data: children come from homes, and they learn in classrooms, which exist within schools, which exist within communities and LAs, which then exist within nations. MLM enables our exploration of these nested, multiple layers of influences. It reflects the nature and contexts of schooling and allows us more accurately to model different sources of potential influences on students and then to relate the structure of education and educational practice to research.

An example of the utility and the explanatory power of this statistical method can be seen in studies about the effects of homework upon achievement, where the use of MLM and of its multiple levels allows us to explore the negative relationship at the individual pupil level and the positive relationship at the school level (Detmers, Trautwein, & Lüdtke, 2009), and studies by D'Haenens Van Damme,

and Onghena (2010) which demonstrate differences in individual- and school-level factor structures.

The usefulness of MLM is further demonstrated by its rapid spread across many different scientific fields. It became popular in many different countries in the early 1980s onwards because it allowed researchers to simultaneously explore individual- and group-level influences. It is now very extensively used in disciplines such as demography (e.g. Sacco & Schmidt, 2005), biology (e.g. McMahon & Diez, 2007), medicine (e.g. Diez-Roux, 2000), and non-educational general social science (e.g. Jones, Johnston, & Pattie, 1992). We know the international social science literature, and Gorard would appear to be in a distinct minority – maybe even a minority of one – when it comes to his cavalier dismissal of the value of MLM approaches. In EER, for example, to argue for the dissolution of MLM as a method returns us to a position where the potential influences of all educational factors default to either the school or the class or the individual level, because we would not have the methodology to disentangle the different sources of influence in the data simultaneously (as is possible with MLM). This would set us back at least three decades in terms of our understanding of educational processes, particularly those relating to the importance of teaching and learning at the classroom level. This approach to statistics risks leaving educational research as a scientific backwater with no credibility among quantitative researchers more generally. The development and increasing application of MLM approaches in different disciplines during the last 30 years reflects their power and academic value in addressing interesting and important research questions with models that can address the complexity inherent in many real-world contexts, rather than the more limited focus on evaluating the impact of interventions used in experimental designs. Even here, there is increasing recognition that MLM can play a valuable role in understanding the variation in the effects of interventions (e.g. via multilevel meta-analysis, as Goldstein, Yang, Omar, Turner, and Thompson (2000) demonstrated in their meta-analysis of the effects of class size).

To conclude, ignoring the multilevel structure of data whether of schools or neighbourhoods leads to biased and inflated estimates of individual-level predictors as well as ignoring the realities of social life.

## *A broader critique of the field*

We move on now to the attempt by Gorard to broaden the critique about methodological issues into a general attack on the core tenets of our field.

### *Doubts about school effects*

Gorard (2011b) doubts that schools have been shown to have any effect, calling the residuals or differences that we utilise in EER 'meaningless' (p. 23, line 14, and p. 17, line 23). He believes that the scale of the errors dwarfs the school residuals but his own data and working of that data does not show this to be the case. He also believes that school effectiveness results show volatility over time because of random events and errors, but in fact the opposite is the case and some recent analyses suggest substantial year-on-year stability in effectiveness (Reynolds, Sammons, DeFraine, Townsend, & Van Damme, 2011), which is impressive given that the great majority of EER research usually takes place within highly unstable communities

and rapidly changing school environments (in terms of personnel and pupil populations).

If school effects were just the result of multiple random events and measurement error, how could it be that across the dozen or so countries where EER has mature research communities, there is so much independent agreement on the size of school effects, their scientific properties, the factors responsible for them, and the ways they can be utilised for school improvement? Using a variety of statistical methods, including MLM, structural equation modelling, and old-fashioned 'means-on-means' approaches, the communality of findings is impressive, especially since these EER communities are very varied in terms of their history, educational commitments, source disciplines, and the nature of the educational processes routinely studied (for an overview see Reynolds et al., 2011; Teddlie & Reynolds, 2000). They are not part of some gigantic monolithic enterprise, as Gorard seems to believe.

Moreover, the more sophisticated our methodological and statistical approaches, the larger the school and teacher effects appear to be. For example, when using an approach that more accurately measures growth in learning using nonlinear multilevel growth curve models, Guldemond and Bosker (2009) show that the school-level variance is between 30% and 50%. Using another innovative method pioneered by educational effectiveness researchers, 'regression discontinuity', Luyten et al. (2009) show school-level variance to be over 33%. If we adopted Gorard's position, we would not even attempt to capture these phenomena, never mind explain the findings. Interestingly, in his *BERJ* article, Gorard (2010a) actually commends the use of regression discontinuity in EER as a methodological approach, without apparently recognising that the authors who apply it do so using MLM first to identify the size of overall absolute school effects and the individual school variance in the size of the school effects.

Gorard (2011b) proceeds to dismiss the EER field in a trivial and unscientific fashion: 'How do we know that all of these school effectiveness studies have produced anything of value at all', he asks (p. 23, line 18), and then answers his own question by asking readers to 'put aside the potted plant theory and those bland almost tautological recommendations for school improvement such as that good schools have good teaching or good leaders'. But EER is a field which, despite the occasionally bland uses to which it is has been put (and we acknowledge this and have written and spoken on the issue many times), is regarded by critics and supporters alike as having made a major impact (e.g. Thrupp, 2001; Townsend, 2007), and one which compares very favourably with any other discipline within educational research in terms of the quality of its insights and the rigour of its scholarship. For example, the journal linked to the International Congress of School Effectiveness and Improvement ranks number 59 out of 177 education journals included within the ISI Citation Indices. Moreover, through the careful study of school and classroom processes and the increasing use of mixed methods research designs (Teddlie & Sammons, 2010), it has provided, for example, in England alone a valuable contemporary knowledge base about effective organisational arrangements (Chapman, Muijs, Sammons, Armstrong, & Collins, 2009, 2011), government intervention (Chapman, Mongon et al., 2011), effective classroom practice (Muijs & Reynolds, 2010), and insights which illuminate the role of leadership in influencing student outcomes indirectly through its impact on school processes and teachers' work (e.g. Day, Sammons, & Gu, 2008; Day et al., 2011).

*The issue of health warnings in EER and conflating EER with government*

Gorard contends that 'Policy makers, schools, departments, teachers, inspectors, parents and authorities are being misled by the spurious results of value added calculations' (p. 24, line 8). He additionally argues that CVA was 'developed with the assistance of school effectiveness researchers' (Gorard, 2011b, p. 9, line 9).

We argued in our *RI* paper (Muijs et al., 2011) that health warnings were indeed necessary in our field and that educational effectiveness researchers had 'constantly and consistently advised policy makers and practitioners accordingly'. We wish to restate here that we have been advised against simplistic translations of EER into policy and the use of CVA without any appreciation of its limitations (e.g. Goldstein, 2008; Sammons, 1996), the need for broader ranges of educational outcomes than academic achievement (Chapman & Gunter 2009; Kelly, 2007; Muijs, 2006; Reynolds, 2010; Sammons, 1996), the need to develop our suite of metrics (Kelly, in press), and the need for caution generally in what EER has and has not found (Mortimore et al., 1988; Reynolds et al., 2011; Sammons, 1999; Teddlie & Reynolds, 2000). We have repeatedly emphasised that EER is a relative and retrospective concept which is both outcome-dependent and time-dependent, and that as a consequence, there is a need to study consistency, stability, differential effectiveness, and trends over time for different groups of pupils (Creemers et al., 2010; Luyten & Sammons, 2010; Sammons, 1996). None of these caveats appear to be known to (or if known, are not acknowledged by) Gorard who claims wrongly that we have not warned audiences about methodological issues (Gorard, 2011b, p. 16, para 2).

It is also important to note as a matter of historical fact that EER researchers had little involvement in the development of the original English CVA measure, in marked contrast to Gorard's claims. To whom is Gorard referring? Some of us were members of the former UK government DfES *Value Added Methodological Advisory Group*, but we consistently pointed out the limitations of existing CVA measures at meetings, as Gorard would know if he had been a member. In any case, this group was set up in 2002 *after* the development of CVA in the early 2000s. In our own EER research (e.g. Muijs & Reynolds, 2000; Sylva, Melhuish, Sammons, Siraj-Blatchford, & Taggart, 2010), we use a much wider range of variables in our analyses than the national CVA system and include a wider range of student outcomes than academic achievement with a strong focus on the topic of educational equity (Sammons, 2010).

The main purpose of EER has never been simply to identify individual schools' performance and rank these but rather, as highlighted in Teddlie and Reynolds (2000), to evidence our knowledge about education through analyses that study the nature of the influences of schools, departments, and classes, including addressing topics such as differential effectiveness (for different student groups reflecting our interest in equity), stability (over time), consistency (on different outcomes), and the school and classroom processes that predict outcomes for students. We are interested in cross-level relationships (how schools may influence classroom practice of teachers and provide a better context for effective teaching and learning). We have consistently sought to provide research evidence of relevance – for the policy community of course and the practice community – to the improvement of both policy and practice and enhancing of understanding of the dynamic processes of educational change (Creemers & Kyriakides, 2008).

We also find it rather surprising that Gorard seeks to cast so many doubts on the historical data collected in the NPD in England in order to critique CVA, yet is

apparently happy to use measures from it in his own research on changes in school composition and to use it to draw conclusions about equity differences in educational achievement (Gorard, 2009, 2010b), without including any analysis to take account of the problems of measurement error that he raises in his criticisms of EER and our work when *we* use such data. We suggest that using individual pupil-level FSM to measure aggregate differences in school composition and to draw conclusions about changes in equity in composition as a result of such an aggregate school-level analysis without reference to measurement error issues and associated CIs could also be misleading since there will be measurement error at both the student level and therefore also at the school level in each year that would affect the calculation of trends across years.

While we are aware of the limitations of the government approach to CVA (and have discussed this extensively ourselves), we do think it appropriate to recognise it represented the first attempt by any government to move beyond reliance on publishing only raw school results and the setting of simple benchmarks to measure school performance. The incorporation of CIs into the calculation of estimates of school performance also marked a recognition that the effectiveness measures are estimates subject to error and prevent the use of ranking tables. Of course, the reliance on one or two overall performance measures ignores subject differences and within-school variance that are of contemporary interest and the focus of much EER. Nonetheless, we see CVA as a step forward and, in fairness, note that it has been widely welcomed by practitioners in schools in difficult circumstances who see it as a means of having their quality and outcomes reflected positively in national data for the very first time by its focus on studying pupil progress rather than just their attainment levels.

In short, the reality about us and our colleagues in EER, about CVA, about the NPD, and about our contribution to policy development is more equivocal, complex, balanced, and multi-faceted than the one-sided account employed by Gorard allows for.

### *The need to avoid personal abuse*

We have dealt with the issue of the credibility of the recent critique in terms of the knowledge content. In this final section, we move on to consider issues to do with the manner of the critique. Put simply, was the critique conducted with integrity and ethically?

One of the most disappointing aspects of this debate is the manner in which it has been conducted. We anticipated a discussion, focusing on technical and methodological issues. However, we did not envisage the unprofessional tone, occasionally bordering on the defamatory, which Gorard chose to take. For example, it is claimed we 'Make no good points at all [and we] betray a basic knowledge of statistics, of the datasets they use in their own work and what is justified in social science writing' (Gorard, 2011b, p. 1, line 24), but this is simply not the case. We made accurate technical points which have been discussed and validated by senior statisticians and social statisticians in leading British and overseas university departments, most of which are not working in the field of education.

It is claimed that we have a 'low level of research craft' (p. 24, line 13), that we are 'lazy and frankly quite stupid' (p. 19, line 18), and that we would 'fail an undergraduate assessment of statistics' (Gorard, 2011b, p. 11, line 22). This puerile banter

is not worthy of any academic debate and certainly not worthy of any academic publication.

More seriously, Gorard (2011b) states (twice) that we have 'rigged' data, once on page 23 (line 38) and once on page 18 (line 2). The Free Online Directory defines the word 'rigged' as 'to manipulate dishonestly for personal gain'. The Meriam Webster Dictionary defines it as (a verb) 'to manipulate or control usually by deceptive or dishonest means' and 'to fix in advance for a desired result', and (as a noun) 'a swindle'. These are very serious allegations.

None of us has ever seen in decades of academic life the use of such inappropriate language in an academic publication or a paper that uses language as this one does about existing educational research. Is it any surprise that the broader academic educational practice and policy communities may doubt the value of educational research evidence when researchers indulge in this kind of anti-intellectual personal attack? BERA's (2011) own ethical guidelines require educational researchers 'to protect the integrity and reputation of educational research and not bring it into disrepute by criticising other researchers in a defamatory or unprofessional manner' (p. 10). Gorard has clearly broken these guidelines in the paper uploaded to the BERA Conference website.

Furthermore, we are taken to task in Gorard (2011b) for not interacting with him and sending him papers as and when he demanded. We refused to do this after his barrage of hostile, insulting and offensive emails, which ironically started after we advised him as a courtesy of our intention to respond to his *BERJ* paper and had given him an advance copy! All our publications are listed and widely available. We believe he has defamed us, but in the interests of furthering academic debate between people, we will not escalate this matter or respond in kind to it.

## Conclusions

To summarise, we believe that the field of EER has had some success in improving the prospects of the world's children over the last three decades – in combating the pessimistic belief that 'schools make no difference', in generating a reliable knowledge base about 'what works' for practitioners to use and develop, and in influencing educational practices and policies positively in many countries.

Over the last 10 years, there have naturally been multiple criticisms of our approaches and findings, but these were often valid ones. In many cases, their utility can be seen in school effectiveness and improvement researchers' responses to these criticisms and in an improvement in the quality of the research undertaken in the field itself.

However, Gorard has taken the criticism to another level – of wholesale rejection of the existence of any educational effect at all, of rejection of the core methodological approaches such as MLM which are widely used within social science research internationally, and of portraying educational effectiveness researchers as tools of governments. Many of his criticisms of us have also taken the stance of what can only be called personal abuse.

Critics, of course, can do and say what they want; that is their personal business. But if the critiques are wrong and have the potential to seriously undermine both the continued growth of the field and its continued professional status, then that is our business; that is why we have contributed this article. We very much hope that others – wherever

they may stand in these debates – will give the community of educational researchers their views in this journal and in others.

## Notes

1. For an explanation and summary of the key features of contextual value added, see http://www.education.gov.uk/performancetables/pilotks4_05/aboutcva.shtml (accessed 15 March 2012).
2. For those unfamiliar with CVA as used by government school residuals are added to a figure of 100 after calculation to avoid possible confusion in interpretation of negative residuals. This is ignored in his calculations.

## References


Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society, Series A*, *149*(1), 1–43.

BERA. (2011). *Ethical guidelines for educational research*. British Educational Research Association. Retrieved August 31, 2011, from http://www.bera.ac.uk/files/2011/08/BERA-Ethical-Guidelines-2011.pdf

Bernstein, B. (1968). Education cannot compensate for society. *New Society*, *387,* 344–347.

Chapman, C., Armstrong, P., Harris, A., Muijs, D., Reynolds, D., & Sammons, P. (Eds.). (2012). *School effectiveness and school improvement research: Challenging the orthodoxy?* London: Routledge.

Chapman, C., & Gunter, H.M. (Eds.). (2009). *Radical reforms: Perspectives on an era of change*. London: Routledge.

Chapman, C., Mongon, D., Muijs, D., Williams, J., Pampaka, M., Wakefield, D., & Weiner, S. (2011). *Evaluation of the extra mile*. London: DfE.

Chapman, C., Muijs, D., & MacAllister, J. (2011). *A study of the impact of federation on student outcomes*. Nottingham: National College.

Chapman, C., Muijs, D., Reynolds, D., Sammons, P., & Teddlie, C. (2013). *International handbook of educational effectiveness research*. London: Routledge.

Chapman, C., Muijs, D., Sammons, P., Armstrong, P., & Collins, A. (2009). *The impact of federations on student outcomes*. Nottingham: National College.

Coleman, J.S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfield, F., & York, R. (1966). *Equality of educational opportunity*. Washington, DC: US Government Printing Office.

Connolly, P., Miller, S., & Eakin, A. (2010). *A cluster randomised trial evaluation of the media initiative for children: Respecting difference programme*. Belfast: Centre for Effective Education, University of Belfast. **Q6**

Creemers, B., Kyriakides, L., & Sammons, P. (Eds.). (2010). *Methodological advances in educational effectiveness research*. London: Routledge.

Creemers, B.P.M. (1994). *The effective classroom*. London: Cassell.

Creemers, B.P.M., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. London: Routledge.

Cuban, L. (1983). Effective schools: A friendly but cautionary note. *Phi Delta Kappan*, *64,* 695–696.

Cuttance, P. (1982). Reflections on the Rutter ethos: The professional researchers' response to fifteen thousand hours: Secondary schools and their effects on children. *Urban Education*, *16*(4), 483–492.

Day, C., Sammons, P., & Gu, Q. (2008). Combining qualitative and quantitative methodologies in research on teachers' lives, work, and effectiveness: From integration to synergy. *Educational Researcher*, *37*(6), 330–342.

Day, C., Sammons, P., Leithwood, K., Hopkins, D., Gu, Q., Brown, S., & Ahtaridou, E. (2011). *Successful leadership: Linking learning outcomes*. Milton Keynes: Open University Press.

Day, C., Sammons, P., Stobart, G., Kingston, A., & Gu, Q. (2007). *Teachers matter*. Buckingham: Open University Press. **Q7**

Department of Education and Science. (1967). *Children and their primary schools* (Plowden Report). London: HMSO.

Detmers, S., Trautwein, U., & Lüdtke, O. (2009). The relationship between homework time and achievement is not universal: Evidence from multilevel analyses in 40 countries. *School Effectiveness and School Improvement*, *20*(4), 375–405.

D'Haenens, E., Van Damme, J., & Onghena, P. (2010). Multilevel exploratory factor analysis: Illustrating its surplus value in educational effectiveness research. *School Effectiveness and School Improvement*, *21*(2), 209–235.

Diez-Roux, A.V. (2000). Multilevel analysis in public health research. *Annual Review of Public Health*, *21,* 171–192.

Elliott, J. (1996). School effectiveness research and it's critics: Alternative visions of schooling. *Cambridge Journal of Education*, *26*(2), 199–223.

Ferrao, M.E., & Goldstein, H. (2008). Adjusting for measurement error in value added models: Evidence from Portugal. *Quality and Quantity*, *43*(6), 951–963.

Fitz-Gibbon, C. (1992). School effects at a level: Genesis of an education system. In D. Reynolds & P. Cuttance (Eds.), *School effectiveness: Research, policy and practice* (Chap. 5, pp. 96–120). London: Cassell. **Q8**

Fitz-Gibbon, C. (1996). *Monitoring education: Indicators, quality and effectiveness*. London: Cassell.

Fitz-Gibbon, C. (1997). *The value added national project final report*. London: SCAA.

Goldstein, H. (1980). Critical notice: Fifteen thousand hours by Rutter et al. *Journal of Child Psychology*, *21*(4), 364–366.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Arnold.

Goldstein, H. (2008). Evidence and education policy – some reflections and allegations. *Cambridge Journal of Education*, *38,* 393–400.

Goldstein, H., Kounali, D., & Robinson, A. (2008). Modelling measurement errors and category misclassifications in multilevel models. *Statistical Modelling*, *8*(3), 243–261.

Goldstein, H., & Noden, P. (2004). A response to Gorard on social segregation. *Research Papers in Education*, *30*(3), 441–444.

Goldstein, H., & Spiegelhalter, D.J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society*, *159A,* 385–443.

Goldstein, H., Yang, M., Omar, R., Turner, R., & Thompson, S.G. (2000). Meta analysis using multilevel models with an application to the study of class size effects. *Journal of the Royal Statistical Society*, *49C,* 399–412.

Good, T.L., & Brophy, J.E. (1986). School effects. In M. Wittrock (Ed.), *Third handbook of research on teaching* (pp. 570–602). New York, NY: Macmillan.

Gorard, S. (2007). The dubious benefits of multi-level modeling. *International Journal of Research and Method in Education*, *30*(2), 221–236.

Gorard, S. (2009). Does the index of segregation matter? The composition of secondary schools in England since 1996. *British Educational Research Journal*, *35*(4), 639–652.

Gorard, S. (2010a). Serious doubts about school effectiveness. *British Educational Research Journal*, *36*(5), 745–766.

Gorard, S. (2010b). Education can compensate for society – a bit. *British Journal of Studies in Education*, *58*(1), 47–65.

Gorard, S. (2011a). Doubts about school effectiveness exacerbated – by attempted justification. *Research Intelligence*, *114,* 26. **Q8**

Gorard, S. (2011b). *Comments on 'The value of educational effectiveness research'*. Retrieved November 18, 2011, from http://beraconference.co.uk/programme-at-a-glance/

Gray, J. (1981). A competitive edge: Examination results and the probable limits of secondary school effectiveness. *Educational Review*, *33*(1), 25–35.

Guldemond, H., & Bosker, R. (2009). School effects on students' progress – a dynamic perspective. *School Effectiveness and School Improvement*, *20*(2), 255–268.

Harris, A., Jamieson, I., & Russ, J. (1995). A study of effective departments in secondary schools. *School Organisation*, *15*(3), 283–299.

Jencks, C.S., Smith, M., Ackland, H., Bane, M.J., Cohen, D., Ginter, H., Heyns, B., & Michelson, S. (1972). *Inequality: A reassessment of the effect of the family and schooling in America*. New York, NY: Basic Books.

Jones, K., Johnston, R.J., & Pattie, C.J. (1992). People, places and regions: Exploring the use of multi-level modelling in the analysis of electoral data. *British Journal of Political Science*, *22,* 242–380.

Kelly, A. (2007). *School choice and student well-being: Opportunity and capability in education: Reviewing the research and adapting Sen's theory of capability to school choice*. London: Palgrave Macmillan.

Kelly, A. (in press). Measuring 'equity' and 'equitability' in school effectiveness research. *British Educational Research Journal.* DOI: 10.1080/01411926.2011.605874 **Q9**

Luyten, H., & Sammons, P. (2010). Multilevel modelling. In B.P.M. Creemers, L. Kyriakides, & P. Sammons (Eds.), *Methodological advances in educational effectiveness research.* (Chap. 11, pp. 246–276). London: Routledge. **Q8**

Luyten, H., Tymms, P., & Jones, P. (2009). Assessing school effects without controlling for prior achievement. *School Effectiveness and School Improvement*, *20*(2), 145–165.

McMahon, S.M., & Diez, J.M. (2007). Scales of association: Hierarchical linear models and the measurement of ecological systems. *Ecology Letters*, *10,* 437–452.

Miller, S., Connolly, P., & Maguire, L. (2011). The effects of a volunteer mentoring programme on reading outcomes among eight- to nine-year-old children: A follow up randomized controlled trial. *Journal of Early Childhood Research*. doi:10.1177/1476718X11407989. **Q10**

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School matters: The junior years*. Wells: Open Books (republished 1995 London: Paul Chapman).

Mortimore, P., Sammons, P., & Thomas, S. (1994). School effectiveness and value added measures. *Assessment in Education: Principles, Policy and Practice*, *1*(3), 315–332.

Muijs, D. (2006). New directions for school effectiveness research: Towards school effectiveness without schools? *Journal of Educational Change*, *7*(3), 141–160.

Muijs, D., Kelly, T., Sammons, P., Reynolds, D., & Chapman, C. (2011). The value of educational effectiveness research – a response to recent criticism. *Research Intelligence*, *114,* 24–25.

Muijs, D., & Reynolds, D. (2000). School effectiveness and teacher effectiveness: Some preliminary findings from the evaluation of the mathematics enhancement programme. *School Effectiveness and School Improvement*, *11*(3), 247–263.

Muijs, D., & Reynolds, D. (2010). *Effective teaching, evidence and practice* (3rd ed.). London: Sage.

Muijs, D., Armstrong, P., & Chapman, C. (2011). *An independent evaluation of the impact of teach first*. London: Goldman Sachs.

Nuttall, D.L., Goldstein, H., Prosser, R., & Rasbash, J. (1989). Differential school effectiveness. *International Journal of Educational Research*, *13,* 769–776.

Plewis, I., & Fielding, A. (2003). What is multilevel modelling for? A critical response to Gorard. *British Journal of Educational Studies*, *51*(4), 408–418.

Purkey, S., & Smith, M. (1983). Effective schools: A review. *Elementary School Journal*, *83,* 427–452.

Reynolds, D. (1976). The delinquent school. In M. Hammersley & P. Woods (Eds.), *The process of schooling* (pp. 217–229). London: Routledge and Kegan Paul. **Q8**

Reynolds, D. (2007). *Schools learning from their best: The within school variation project*. Nottingham: NCSL.

Reynolds, D. (2010). *Failure free schooling*. London: Routledge.

Reynolds, D., Sammons, P., DeFraine, B., Townsend, T., & Van Damme, J. (2011, January). *Educational effectiveness research: A state of the art review*. Paper given to the Annual Conference of the International Congress of School Effectiveness and Improvement, Cyprus.

Reynolds, D., & Sullivan, M. (1981). The effects of school: A radical faith restated. In B. Gilham (Ed.), *Problem behaviour in secondary school*. London: Croom Helm.

Reynolds, D., & Teddlie, C. (2001). Reflections on the critics, and beyond them. *School Effectiveness and School Improvement*, *12*(3), 99–114.

Rutter, M., Maughan, B., Mortimore, P., & Ouston, J. (1979). *Fifteen thousand hours: Secondary schools and their effects on children*. London: Open Books.

Sacco, J.M., & Schmidt, N. (2005). A dynamic multilevel model of demographic diversity and misfit effects. *Journal of Applied Psychology*, *90,* 203–231.

Sammons, P. (1996). Complexities in judging school effectiveness. *Educational Research and Evaluation*, *2*(1), 113–149.
Sammons, P. (1999). *School effectiveness: Coming of age in the 21st century*. Lisse: Swets and Zeitlinger.
Sammons, P. (2010). Equity and educational effectiveness. In P. Peterson, E. Baker, & B. McGaw (Eds.), *The international encyclopaedia of education* (Vol. 5, pp. 51–57). Oxford: Elsevier. **Q11**
Sammons, P., Hillman, J., & Mortimore, P. (1995). *Key characteristics of effective schools: A review of school effectiveness research*. London: Ofsted.
Sammons, P., Nuttall, D., & Cuttance, P. (1993). Differential school effectiveness: Results from a reanalysis of the Inner London Education Authority's junior school project data. *British Educational Research Journal*, *19*(4), 381–405.
Sammons, P., Thomas, S., & Mortimore, P. (1997). *Forging links: Effective schools and effective departments*. London: Chapman.
Slee, R., & Weiner, G. (2001). Education reform and reconstruction as a challenge to research genres: Reconsidering school effectiveness research and inclusive schooling. *School Effectiveness and School Improvement*, *12,* 83–98.
Slee, R., Weiner, G., & Tomlinson, S. (1998). *School effectiveness for whom? Challenges to the school effectiveness and school improvement movements*. London: Falmer Press.
Smith, D.J., & Tomlinson, S. (1989). *The school effect: A study of multi-racial comprehensives*. London: Policy Studies Institute.
Strand, S. (2010). Do some schools narrow the gap? Differential school effectiveness by ethnicity, gender, poverty and prior achievement. *School Effectiveness and School Improvement*, *21,* 289–314.
Sylva, K., Melhuish, E., Sammons, P., Siraj-Blatchford, I., & Taggart, B. (2010). *Early childhood matters*. London: Routledge.
Teddlie, C. (2009). The legacy of the school effectiveness research tradition. *Second International Handbook of Educational Change*, *23*(3), 523–554.
Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research*. London: Falmer Press.
Teddlie, C., & Reynolds, D. (2001). Countering the critics: Responses to recent criticisms of school effectiveness research. *School Effectiveness and School Improvement*, *12,* 41–82.
Teddlie, C., & Sammons, P. (2010). Applications of mixed methods to the field of educational effectiveness research. In B.P.M. Creemers, L. Kyriakides, & P. Sammons (Eds.), *Methodological advances in educational effectiveness research* (Chap. 7, pp. 115–152). London: Routledge. **Q8**
Thrupp, M. (1999). *Schools making a difference: Let's be realistic*. Buckingham: Open University Press.
Thrupp, M. (2001). Sociological and political concerns about school effectiveness research: Time for a new research agenda. *School Effectiveness and School Improvement*, *12,* 7–40.
Townsend, T. (2001). Satan or saviour? An analysis of two decades of school effectiveness research. *School Effectiveness and School Improvement*, *12,* 115–129.
Townsend, T. (Ed.). (2007). *The international handbook of school effectiveness and improvement*. Dordrecht: Springer.
Van Damme, J., Van Landeghem, G., Onghena, P., De Fraine, B., & Opdenakker, M.-C. (2001, September). *The effect of schools and classes upon well-being and other non-cognitive outcomes*. Paper presented at the European Conference on Educational Research, Lille.
Van de Gaer, E., De Fraine, B., Van Damme, J., De Munter, A., & Onghena, P. (2009). School effects on the development of motivation toward learning tasks and the development of academic self-concept in secondary school. *School Effectiveness and School Improvement*, *20,* 235–253.
Willmott, R. (1999). School effectiveness research: An ideological commitment? *Journal of Philosophy of Education*, *33*(2), 253–268.
Woodhouse, G., Yang, M., Goldstein, H., & Rasbash, J. (1996). Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society, Series A*, *159*(2), 201–212.
Wrigley, T. (2004). 'School effectiveness': The problem of reductionism. *British Educational Research Journal*, *30*(2), 227–244.