Education Economics

ISSN: 0964-5292 (Print) 1469-5782 (Online) Journal homepage: http://www.tandfonline.com/loi/cede20

Revisiting fixed- and random-effects models: some considerations for policy-relevant education research

Paul Clarke, Claire Crawford, Fiona Steele & Anna Vignoles

To cite this article: Paul Clarke, Claire Crawford, Fiona Steele & Anna Vignoles (2015) Revisiting fixed- and random-effects models: some considerations for policy-relevant education research, Education Economics, 23:3, 259-277, DOI: 10.1080/09645292.2013.855705

To link to this article: http://dx.doi.org/10.1080/09645292.2013.855705

Education Economics, 2015
Vol. 23, No. 3, 259–277, http://dx.doi.org/10.1080/09645292.2013.855705

Routledge
Taylor & Francis Group

Revisiting fixed- and random-effects models: some considerations for policy-relevant education research

Paul Clarke[a], Claire Crawford[b], Fiona Steele[c] and Anna Vignoles[d*]

[a]*ISER, University of Essex, Colchester, UK;* [b]*Institute for Fiscal Studies, London, UK;* [c]*Department of Statistics, London School of Economics, London, UK;* [d]*Faculty of Education, University of Cambridge, Cambridge, UK*

(*Received 29 May 2012; accepted 10 October 2013*)

The use of fixed (FE) and random effects (RE) in two-level hierarchical linear regression is discussed in the context of education research. We compare the robustness of FE models with the modelling flexibility and potential efficiency of those from RE models. We argue that the two should be seen as complementary approaches. We then compare both modelling approaches in our empirical examples. Results suggest a negative effect of special educational needs (SEN) status on educational attainment, with selection into SEN status largely driven by pupil level rather than school-level factors.

Keywords: fixed effects; random effects; hierarchical linear regression

1. Introduction

Hierarchical regression models allow for data structures where individual study units are nested within higher level units, such as pupils within schools and persons within households. These structures often lead to clustering when, for example, the achievements of pupils in the same school are clustered due to the influence of unmeasured school characteristics like ethos and teacher quality. A key choice to be made when fitting hierarchical models is whether to treat the higher level terms as fixed or random. In multidisciplinary research areas such as education, the discipline of the researcher often influences the nature of the research question being addressed, which in turn determines the choice of modelling approach. Economists often focus on estimating the 'causal' effects of predictor variables on outcomes, for which they are more likely to use fixed-effects (FE) models to remove sources of higher level selection. In other quantitative disciplines, the primary aim is to explain and estimate the contribution of higher level units to the variation in individuals' outcomes, such as in the extensive school effectiveness literature, for which RE models are regularly used.

Much is known about the relative strengths of the FE and RE approaches, and the issue has already received considerable attention in the literatures from econometrics (e.g. Wooldridge 2002; Cameron and Trivedi 2005) and statistics (e.g. Snijders and Bosker 2011; Gelman and Hill 2007). However, not all applied social researchers are familiar with this work, and there remains some confusion over which approach to use. The purpose of this article is to review this work in a non-technical manner for

*Corresponding author. Email: av404@cam.ac.uk

the general reader and to promote an analysis strategy in which these approaches are viewed as complementary rather than competing. By doing so, we make a contribution to multidisciplinary understanding at a time of increasing interest in the use of observational studies to assess current policy and influence future policy direction.

For illustration, we use two topical empirical examples from the education field on the impact of special educational needs (SEN) interventions on pupils' educational attainment. These examples are apposite given that education research is a multidisciplinary field in which controversies have arisen about the use of FE and RE (multilevel) models, but we hope it is clear how our analysis is relevant to all policy-relevant social science research. The remainder of this article is set out as follows. In Section 2, we review the key features of FE and RE models and in Section 3, we discuss the impact of selection on model choice. In Section 4, we present the results of two illustrative examples on the impact of SEN interventions on pupil achievement. Finally, in Section 5 we make some concluding remarks. Note that we focus here on the problems caused by selection, and so sidestep the related problem of covariate measurement error; the interested reader is referred elsewhere for further details about this important issue (e.g. Woodhouse et al. 1996; Ebbes, Bockenholt, and Wedel 2004).

2. Hierarchical models

Individuals are often part of a hierarchical structure in which they are members of higher level units. If individuals are grouped in higher level units, such as schools, this can lead to statistical clustering if belonging to the school in this case has an independent influence on individuals' outcomes. For example, the achievements of pupils in the same school are likely to be clustered due to the influence of unmeasured school characteristics, such as school ethos, which influence their academic progress after they join the school.

Without loss of generality, we refer to the higher level units as 'schools' throughout the following discussion. A two-level hierarchical model of individual pupils within schools sets individuals at level 1 and schools at level 2. Letting y_{ij} be the outcome of individual i in school j ($i = 1, \ldots, n_j; j = 1, \ldots, J$), the hierarchical regression model can be written as

$$y_{ij} = \beta_0 + x_{1ij}\beta_1 + \cdots + x_{pij}\beta_p + u_j + e_{ij} = \beta_0 + \boldsymbol{x}_{ij}'\boldsymbol{\beta} + u_j + e_{ij}, \tag{1}$$

where β_0 is the intercept term, $\boldsymbol{x}_{ij}$ is a vector representing the p individual-level covariates/regressors and $\boldsymbol{\beta}$ is the corresponding vector of regression coefficients, or 'slopes', for each of these covariates. As we will see, the intercept term is not strictly necessary for FE models, but we keep it in place for now. We start by viewing Equation (1) purely as a model of the association between the outcome and the covariates, and so follow the convention used in the statistical literature on multilevel modelling by referring to e_{ij} as the individual-level *residual* (e.g. Goldstein 2010). The population average of the individual-level residuals e_{ij} is zero.

The definition of the school-level term u_j depends on whether we treat it as a 'fixed' or a 'random' effect, but before we consider the two approaches in detail, we start by considering some modelling assumptions common to both. Specifically, these assumptions concern the individual-level residual. The simplest assumption is that the individual-level residuals are *homoskedastic*, that is, the subpopulations defined by $\boldsymbol{x}_{ij}$ from

which the individuals' residuals are drawn all have the same variance so that $\text{var}(e_{ij}|\boldsymbol{x}_{ij}) = \text{var}(e_{ij}) = \sigma_e^2$. We make this assumption to simplify what follows, but it is not essential and is easily relaxed to allow *heteroskedastic* individual-level residuals where the residual variance can depend on the covariates (e.g. Wooldridge 2002, Ch. 10; Goldstein 2010, Ch. 3).

The next assumption is that the individual-level residuals are normally distributed. For homoskedastic residuals, this assumption is often expressed as $e_{ij} \sim$ i.i.d. $N(0, \sigma_e^2)$, that is, the residuals are all drawn independently from the same normally distributed population. However, despite often being thought of as essential, normality is not required for estimates of the regression coefficients of linear models like Equation (1) to be unbiased and consistent for $\boldsymbol{\beta}$; in fact, even the standard errors of these estimates will be approximately correct when based on large samples (of both individuals and schools) if the individual-level residuals are non-normal.[1]

The essential assumption for policy-relevant inference is that the individual-level residuals are independent of $\boldsymbol{x}_{ij}$. In the economics literature, this is known as the 'exogeneity' or 'orthogonality' assumption, but we refer to it as the *regression assumption* to indicate its importance for both FE and RE regression modelling. The regression assumption is crucial for ensuring that the regression coefficients are not merely measures of association, but have a policy-relevant, or causal, interpretation. Crucially, the regression assumption **cannot** be guaranteed to hold when analysing data from observational studies. We expand on what is meant by 'causal' and 'policy-relevant' in Section 3 and return to this crucial issue again further on.

2.1 *RE models*

RE models are also known as *multilevel* or *mixed-effects models*. In the multilevel modelling literature, the school-level terms u_j are referred to as school-level residuals and are treated as school-level equivalents of the individual-level residuals so that $u_j \sim$ i.i.d. $N(0, \sigma_u^2)$. It is often assumed by researchers that the school-level residuals are homoskedastic and normally distributed but, as with individual-level residuals, the normality assumption is not crucial if the main concern is estimating $\boldsymbol{\beta}$. The RE version of model (1) is also known as a 'random intercepts' model because $\beta_{0j} = \beta_0 + u_j$ is interpreted as the school-specific intercept for school j and $\beta_{0j} \sim$ i.i.d. $N(\beta_0, \sigma_u^2)$ is random. Under a random intercepts model, the total residual variance can be partitioned into two components: the between-school variance σ_u^2 and the within-school (between-individual) variance σ_e^2.

RE models are popular in many research disciplines and particularly in education. A prime motivation for researchers using RE models is that the clustering of individuals within schools is treated as a feature of scientific interest in its own right, which leads to three important advantages: (1) relationships between the characteristics of the school-level unit and the individual-level outcome of interest can be examined by including school-level covariates; (2) *shrunken* estimates of the school-level residuals can be straightforwardly obtained using standard software; and (3) RE models can be extended to also include 'random coefficients' that allow the effects of predictor variables to vary between higher level units. In general, a major advantage of the RE model is of course its ability to identify the effects of higher level variables on the outcome of interest; in the context of education, this is something that is of particular importance when considering the role of school characteristics.

Shrinkage estimators have long been established as the best class of estimator for predictions (e.g. Efron and Morris 1973). The shrunken estimate of a school residual is

$$\hat{u}_{j,\mathrm{RE}} = \left(\frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_j}\right)\left(\bar{y}_j - \hat{\beta}_{0,\mathrm{RE}} - \bar{\boldsymbol{x}}_j'\hat{\boldsymbol{\beta}}_{\mathrm{RE}}\right), \tag{2}$$

where $\bar{y}_j$ and $\bar{\boldsymbol{x}}_j$ are school means of y_{ij} and $\boldsymbol{x}_{ij}$, respectively, n_j is the number of individuals in school j and $\hat{\beta}_{0,\mathrm{RE}}$, $\hat{\boldsymbol{\beta}}_{\mathrm{RE}}$, $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$ are estimates of the intercept, slopes, school-level variance and individual-level variance, respectively, based on the RE model. In many studies, the school-level residuals are of substantive interest, and so it is crucially important that researchers can interpret these estimates with confidence. Shrinkage estimators do this by down-weighting 'rogue' estimates of school residuals based on only a few pupils within a school, which may be particularly large or small due to sampling variation (e.g. Aitkin and Longford 1986; Goldstein 1997).

Random coefficients, or random slopes, models are an extension of Equation (1) to allow the effects of individual-level covariates to vary across schools. For example, a random coefficients model for covariate x_{1ij} involves specifying an additional random effect u_{1j} such that the coefficient of x_{1ij} is $\beta_{1j} = \beta_1 + u_{1j}$. The j subscript on β_{1j} indicates that the effect of covariate x_{1ij} varies between schools. These models are widely used in social research (e.g. Nuttall et al. 1989; Sammons, Nuttal, and Cuttance 1993), and in our illustrative examples we use random coefficients to model between-school heterogeneity in the effect of special education needs interventions on pupil achievement (see Section 4).

Unfortunately, the flexibility of RE models comes at a cost, which is that we must assume that u_j is uncorrelated with any of the covariates $\boldsymbol{x}_{ij}$ in our model.[2] In the economics literature, this is referred to as the *RE assumption*. If the RE assumption fails, then RE models cannot be used for policy-relevant causal inference. Using models of pupil achievement as an example, this assumption implies that unobserved characteristics of the school u_j that influence achievement, such as school ethos, are uncorrelated with the individual-level and school-level characteristics included in the model (e.g. whether a pupil is classified as having SEN). We discuss the RE assumption again in Section 3.

2.2 *FE models*

There are two equivalent FE approaches to fitting hierarchical model (1). One involves including the schools as covariates and the other 'differences out' the school-level terms; we focus on the first of these approaches. One way of specifying a FE model is as

$$y_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \sum_{k=1}^{J} d_{ik}u_k + e_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + u_j + e_{ij}. \tag{3}$$

It can be seen that the intercept β_0 has been dropped from the model (or set to zero) and J dummy variables for school membership have been included along with the usual covariates $\boldsymbol{x}_{ij}$. The dummy variable d_{ik} equals 1 if individual i is in school k and equals 0 otherwise; hence, $d_{ij} = 1$ in the equation above because individual i is in school j. By

specifying the model in this way, the coefficients of the dummy variables u_j correspond to school-specific intercepts, which replace the overall intercept β_0 and are the fixed-effect equivalents of the RE model's random intercepts β_{0j}. There are, however, other ways to parameterise FE models. For example, we could also choose one school to be the reference school, and specify a regression model in which β_0 is included along with $J-1$ dummy variables for every school bar the reference. But model (3) is the easiest to interpret and so we use it for the purposes of our discussion.

The FE models require no parametric distributional assumptions about u_j, but unlike RE models cannot be used to estimate the effects of school-level covariates. The school residuals can be estimated by $\hat{u}_{j,\mathrm{FE}} = \bar{y}_j - \bar{\boldsymbol{x}}_j' \hat{\boldsymbol{\beta}}_{\mathrm{FE}}$, where $\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$ is obtained from fitting (3). In contrast to the shrunken residuals from RE models, these estimates may be unreliable *when the school size is small* or the *within-school variance is large relative to the between-school variance*. Small sample sizes are an issue because we may obtain very large positive or negative values of $\hat{u}_{j,\mathrm{FE}}$ simply by chance, which makes $\hat{u}_{j,\mathrm{FE}}$ an inaccurate estimate of the true u_j and makes it difficult to compare estimates of school residuals based on different numbers of pupils. Where analyses are based on population administrative data, these limitations may be inconsequential, but in analyses using data on small numbers of individuals, these estimates can be poor because sampling variability will lead to some estimates being extremely small or extremely large relative to the true effect (e.g. Goldstein 1997).

2.3 *The Hausman test*

A full technical discussion of the relationship between the RE and FE estimators can be found in Wooldridge (2002, Sec. 10.7). Of note is that the two estimators are approximately the same in the following scenarios: (a) if the between-school variation is small (in which case both estimators behave like a simple OLS regression model without school residuals/effects); (b) if the number of individuals within the school is large; (c) if the between-school variation is large *relative* to the within-school variation, which may be likely with a high degree of sorting into schools; or (d), trivially, if the RE assumption holds.

Outside of these scenarios, concern about the RE assumption failing often leads researchers to try to formally test whether it holds. In the economics literature, this is conventionally done using the Hausman test (e.g. Wooldridge 2002, Sec. 10.7.3). The null hypothesis of the Hausman test is that both the FE and RE estimators are targeting the same value of $\boldsymbol{\beta}$. For a particular covariate k, the Hausman test statistic is

$$h_k = \frac{\hat{\beta}_k^{\mathrm{FE}} - \hat{\beta}_k^{\mathrm{RE}}}{\sqrt{\mathrm{var}(\hat{\beta}_k^{\mathrm{FE}}) - \mathrm{var}(\hat{\beta}_k^{\mathrm{RE}})}}, \tag{4}$$

which is approximately normally distributed (with a mean of zero and variance of one) in large samples, provided that the model's variance structure has been correctly specified as either homoskedastic or heteroskedastic. Rejecting the null hypothesis is often used to indicate failure of the RE assumption, but there are caveats to bear in mind about this interpretation, of which we highlight two. The first is common to all significance tests when applied to large samples, namely, a small difference between the FE

and RE estimates can lead to rejection of the null hypothesis even when the difference between the estimates is substantively insignificant. Second, if the regression assumption fails then the test is biased because both models are mis-specified and the asymptotic distribution of Equation (4) is incorrect (e.g. Wooldridge 2002, Sec. 10.7.3; simulation results available on request from the authors). Both of these points indicate that, while the Hausman test has a role to play in comparing the estimates obtained from FE and RE models, it does not necessarily provide the definitive answer about which should be preferred. The reader is referred to a more extensive discussion of issues arising with the Hausman test by Fielding (2004) and Snijders and Berkhof (2008).

3. Policy-relevant inference and selection

We now clarify what is meant by 'policy-relevant' causal inference. Given the aims of this article, we do not attempt to provide a formal treatment of the use of hierarchical models for causal inference, for which we refer the reader elsewhere (e.g. Hong and Raudenbush 2006). Instead, we focus on providing an informal treatment to give some insight into the main issues.

Causal inference concerns causal parameters such as average treatment effects. An average treatment effect is defined as the average difference (over all individuals) between the outcomes of the individual in two different scenarios: one with and one without the treatment with everything else held fixed (that is, the **only** difference between scenarios is the treatment). It is a powerful indicator of what would happen if a policy based on this treatment is implemented more widely in the target population. The 'treatment' is SEN status in our example below. Using data from a perfectly conducted randomised controlled experiment in which SEN is assigned at random, the average treatment effect can be estimated simply by the difference in outcomes between those given the treatment and those untreated, because random allocation of SEN is done irrespectively of each individual's characteristics and school. In observational studies, however, where individuals choose their schools non-randomly, and schools may vary in the way that treatments are allocated (and possibly in the effectiveness or type of treatment used), adjustments for selection must be made.

Consider a simple hierarchical model of individual pupils nested within schools in which the treatment of interest is the only covariate:

$$y_{ij} = \beta_0 + \text{Treat}_{ij}\tau + u_j + e_{ij}, \tag{5}$$

where Treat_{ij} is a binary indicator of the 'treatment' received by individual i in school j and our target parameter τ is the average effect of the intervention treatment. (Recall that the FE equivalent of Equation (5) as described in model (3) includes dummy variables for each school in the model and constrains $\beta_0 = 0$.) If we fitted this model (RE or FE) to data from a randomised experiment then the coefficient of Treat_{ij} equals the average treatment effect. However, if the data come from an observational study, then we must adjust for non-random selection.

First, we must account for individual-level characteristics that are a) associated (positively or negatively) with pupils being given the treatment **and** b) associated with the outcome. Unless this is done then the **regression assumption** introduced in Section 2 will fail, and the coefficient of Treat_{ij} will be a biased estimate of the average treatment effect, regardless of whether FE or RE models are used. These

variables must be included in the model through individual-level covariates so that Equation (5) becomes

$$y_{ij} = \beta_0 + \text{Treat}_{ij}\tau + \boldsymbol{x}'_{ij}\boldsymbol{\beta}_1 + u_j + e_{ij}, \tag{6}$$

where $\boldsymbol{x}_{ij}$ represents the covariates included in the model and $\boldsymbol{\beta}_1$ is the regression coefficients of these covariates. It should be noted that adjustments for non-random treatment selection can be made using regression models, but these depend on the prior knowledge of the researcher about which drivers of selection are most strongly correlated with the outcome. The extent to which selection can be accounted for also depends on the richness of the available data: an investigator may know from previous work that having a sibling at a particular school is a strong driver of school selection and related to attainment; but if no information about siblings is available in the data-set then this factor cannot be adjusted for. Moreover, it is important to recognise that adjustments using regression models may be driven by implicit but unverifiable modelling assumptions (e.g. Blundell, Dearden, and Sianesi 2005; Gelman and Hill 2007, chaps. 10–11).

Second, we must account for school-level factors that are correlated with SEN assignment status **and** individual outcomes. In theory, the (weighted) sum of all the school-level factors omitted from Equation (6) add up to u_j so that we can write the school effect as

$$u_j = \boldsymbol{w}'_j\boldsymbol{\delta} + \boldsymbol{z}'_j\boldsymbol{\lambda}, \tag{7}$$

where $\boldsymbol{w}_j$ represents all of the omitted school-level variables **correlated** with Treat_{ij}, $\boldsymbol{z}_j$ represents all of the omitted school-level variables **uncorrelated** with the covariates and $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ are unknown parameters.[3] It is because of the correlated school-level factors $\boldsymbol{w}_j$ that the RE assumption can fail. For example, take two pupils with the same individual-level characteristics in two schools, A and B, where school A is much less likely to provide for and be selected by children with SEN because it has a strict academic ethos. In this case, the comparison of treated and untreated pupils by the RE model confounds the effect of SEN with the effect of academic ethos.

The FE approach is often chosen by default to avoid this problem because the RE assumption is presumed to be unrealistic. However, we argue that this decision is often made too hastily and ignores that school-level covariates can be included in RE models. While the RE assumption will never hold exactly, by including the school-level $\boldsymbol{w}_j$ hypothesised to contribute the most to u_j in Equation (7), we can seek to ameliorate the impact of its failure. To do this, we must have some knowledge of the school-level factors associated with treatment selection and a data-set containing variables that measure these factors. The RE estimates can then be compared to those obtained using the equivalent FE model to assess whether school-level selection has been satisfactorily controlled for. After adjustment, the school residuals can be straightforwardly interpreted as the effect of the uncorrelated school-level variables $\boldsymbol{z}_j$.

An alternative approach from the RE literature is to include school-level means of the covariates. We can write this model as

$$y_{ij} = \alpha_0 + \text{Treat}_{ij}\tau + \boldsymbol{x}'_{ij}\boldsymbol{\beta}_1 + \text{PTreat}_j\alpha_1 + \bar{\boldsymbol{x}}'_j\boldsymbol{\alpha}_2 + u^*_j + e_{ij}, \tag{8}$$

where PTreat_j is the proportion treated in school j (that is, the school-level mean of Treat_{ij}), $\bar{\boldsymbol{x}}_j$ is the school-level mean of $\boldsymbol{x}_{ij}$, $u^*_j \neq u_j$ is a new random effect, α is a

new intercept term and α_1 and $\boldsymbol{\alpha}_2$ are the 'contextual effects' of Treat_{ij} and covariate(s) $\boldsymbol{x}_{ij}$, respectively (e.g. Skrondal and Rabe-Hesketh 2004; Snijders and Berkhof 2008). Even if the RE assumption fails for Equation (6), the estimate of τ based on Equation (8) will equal the average treatment effect.

The simplicity and robustness of this approach make it appealing, but there are some limitations if the RE and the contextual effects are also of interest. The first limitation is that, generally, the u_j^* cannot be usefully interpreted. The omitted $\boldsymbol{w}_j$ are not simply aggregates of the individuals' characteristics within the school, but genuinely school-level characteristics like ethos; model (8) then corrects for bias provided that the slope of a (hypothetical) linear regression of the omitted w_j on the school-level means is non-zero – *whether or not the true relationship between these variables is linear*. However, if the true relationship is non-linear then u_j^* cannot be interpreted straightforwardly as the effect of the uncorrelated school-level variable(s) $\boldsymbol{z}_j$, because the effect of $\boldsymbol{z}_j$ will be confounded with non-linear effects of the school-level means.[4] The second limitation is that estimates of the contextual effects will be biased if the school-level means are not the true/census means but estimated based on the sample data (e.g. Grilli and Rampichini 2011). The equivalence of FE and RE for models including all the school means was first noted by Mundlak (1978); this close connection is confirmed by the fact that testing equality of the contextual effects with zero is equivalent to the Hausman test (e.g. Snijders and Berkhof 2008).

It is important to point out here that the robustness of FE is not unlimited. Crucially, if treatment selection is most strongly associated with individual-level factors rather than school-level ones, and these are not adequately controlled for in the model, then the advantages of the FE approach are lost: both the RE and FE models will give biased results because the regression assumption fails. The importance of individual-level selection in educational research is illustrated by Burgess et al. (2011), who study the factors affecting the school choices made by parents in England. The dominant factors affecting their decisions were found to be: (a) proximity to school; (b) a sibling attends the same school; (c) family members or friends attend the same school; (d) school reputation; and (e) childcare facilities offered. Only (d) and (e) can be usefully viewed as operating at the school level; the others all depend on the parents and residential location and so vary between families/pupils rather than schools. Any (direct or indirect) association between these factors and the outcomes and treatment will lead to bias.

In summary, the key message is that school-level selection does not necessarily preclude the use of RE models to estimate treatment effects, provided that the school-level variables $\boldsymbol{w}_j$ are included in the model. A second message is that individual-level selection (which is influenced by pupil, family and neighbourhood characteristics) is the most important source of bias and both FE and RE models are sensitive to failure of the regression assumption. With regard to school-level selection, it is unnecessary, and probably impossible, to include all the necessary school-level variables, but an achievable objective is to include as many dominant variables as possible, namely, those most strongly associated with both treatment status and the outcome. Clearly, this is more difficult to justify if the selection mechanism is poorly understood and/or the data-set being analysed is limited in terms of its school-level variables. Thus, the general policy we advocate is to use FE and RE to complement each other: the difference between the estimates based on a school-level covariate-adjusted RE model and an FE model can be compared using the Hausman test (recalling its caveats) and by assessing whether the two estimates lead to substantively different

policy conclusions. If the difference is statistically and substantively significant then the treatment-effect estimates based on FE (or RE with school-level means) should be preferred. These issues are explored in the following illustrative examples.

4. Application: analysis of pupil progress in primary school

We now introduce two illustrative and related examples to demonstrate the practical implications of using different modelling approaches in a specific context. Our illustrations are drawn from an analysis of the relationship between pupil progress in primary school (between ages 7 and 11) and SEN status. One of the examples is based on only administrative data, while in the other we make use of rich survey data.

4.1 *Choice of treatment*

Understanding the impact of SEN status on pupil progress is not only a useful exemplar of many of the issues we want to discuss, but also of crucial policy importance given the large numbers of pupils affected and the resources allocated to SEN interventions.[5] We would like to measure the impact of receiving a specific SEN treatment on pupil performance. We conceptualise the data as coming from a (quasi) experiment in which pupils enter schools and are then identified as having SEN (and hence selected into a particular SEN 'treatment') by their parents and teachers. In practice, however, children with very similar needs are not uniformly assigned to the treatment groups across schools and even those who are identified as having SEN receive a range of different interventions that vary across schools and children's needs. To constrain this heterogeneity, we focus on children who are identified as having non-statemented (or less severe) SEN and exclude those with statemented (or more severe) SEN from our analysis.[6] Even in this relatively homogenous group, however, there is considerable treatment heterogeneity. Furthermore, it is not possible to identify the type of intervention received by each pupil using the available data. We only know whether a pupil is identified as having non-statemented SEN or not. Our aim therefore is to estimate the effect of being labelled as having non-statemented SEN, rather than the effect of a particular SEN intervention, on pupils' academic progress.

Our null hypothesis is that a child who has been identified as having SEN will receive additional academic or other support to meet these needs. Hence, compared to otherwise similar children who have not been identified as having SEN, the children with SEN should have higher levels of achievement or, more specifically, make greater academic progress as measured by the gain in their test scores. An alternative hypothesis, however, is that identifying children as having SEN may have a negative impact. First, it may reduce their confidence and lower their expectations along with those of their parents and teachers. Second, there is evidence that children with SEN spend less time interacting with teachers and far more time with teaching assistants (TAs), who are on average less well qualified than teachers. Indeed, research has indicated that pupils supported by TAs make less progress on average than their similarly able peers (Blatchford et al. 2011; Webster and Blatchford 2013). This issue certainly applies to children with statements of SEN who are allocated TA time and may also apply to those children with significant needs who do not have statements but who tend to be supported by TAs. Consequently, there are a number of possible reasons as to why a child identified as having SEN may perform worse (or make slower progress) than an otherwise identical child who has not been formally identified as having SEN.

A hierarchical analysis of the effects of SEN on pupil progress is appropriate because much of the variation in the incidence of SEN is at the school level due to variation in school and local authority policies (Lamb 2010; Ofsted 2010). We might, a priori, be concerned that pupils with SEN are more likely to attend schools that have particular unobserved characteristics, such as a supportive ethos, and that such characteristics may be correlated with pupil achievement. If inadequately captured by the inclusion of observable school characteristics in a RE model, then this would lead to a failure of the RE assumption.

4.2 *Data sources*

For our first example, we use data from the Avon Longitudinal Study of Parents and Children (ALSPAC).[7] The purpose of this example is to illustrate issues about choice of model when the data are very rich, though sample sizes are limited. ALSPAC is a longitudinal survey focusing on the children of around 14,000 pregnant women who were resident in the Avon area of England and whose expected date of delivery fell between 1 April 1991 and 31 December 1992. The ALSPAC cohort members have been surveyed frequently from the time of pregnancy onwards. Information has been collected on a wide range of family background characteristics, and a variety of physical, psychometric and psychological tests (such as IQ and the Strengths and Difficulties Questionnaire)[8] have also been administered. The study further includes information from the teachers and head teachers of ALSPAC cohort children during primary school. Importantly for our purposes, the cohort members have been linked to their administrative data records from the National Pupil Database (see below for further details), which contains test-score information at ages 7 and 11, and limited personal characteristics that include our covariate of interest (SEN status). Table 1 contains full details of the covariates used in this analysis. Our sample comprises 5417 pupils for whom both the age 7 and age 11 test scores and SEN status are observed. We use standardised average test scores at each age.

Our second example relies solely on data from the National Pupil Database. The National Pupil Database is an administrative data-set containing national achievement test scores and a limited set of personal characteristics for all children attending state schools in England. This example has been selected to illustrate the issues arising when using 'sparse' data (in the sense of containing limited information about each individual), albeit with a very large sample size in this case. We use the same cohorts of pupils as those in the ALSPAC sample, namely, those who sat their Key Stage 2 tests at age 11 in 2001–2002, 2002–2003 and 2003–2004. We limit the sample to those who do not have statements of SEN, as in the example above, and those for whom we observe both Key Stage 1 (age 7) and Key Stage 2 test scores. This yields a sample size of over 1.6 million children. The covariates used in this analysis are the same as those outlined in the top panel of Table 1 for the ALSPAC sample.

Table 2 contains descriptive statistics and a comparison between the ALSPAC sample used in our first example and the administrative data on all children in the relevant cohorts attending state schools in England, which is used in our second example.

4.3 *Analysis strategy*

The objective of our analysis is to illustrate the methodological points raised in Section 3. We consider a range of increasingly complex RE and FE models of academic

Table 1. Explanatory variables by type of data source.

Administrative[a]		
Eligible for free school meals (FSM)	Ethnicity (white vs. non-white)	National achievement test scores at ages 7 and 11
SEN status	English as an additional language	Month of birth, gender
Typical longitudinal survey[b]		
Child circumstances around birth		
Birth weight	Mother's age at birth	Mother's marital status
Multiple birth indicator	Number of older siblings	Ever breastfed
Parental characteristics		
Mother's and father's occupational class	Mother's and father's level of education	
Family circumstances during childhood		
No. of younger siblings (at 81 months)		
Mean household income (at 33 and 47 months)		
Ever in financial difficulties (during pregnancy or at 8, 21 or 33 months)		
Ever lived in council or housing association rented accommodation		
Lived in owner-occupied housing since birth		
Child cognitive and behavioural measures		
Child has internal locus of control at age 8	Self-perception of reading ability at age 9	Child likes school at age 8
Child has external locus of control at age 8	Self-perception of maths ability at age 9	Truant age 7 (teacher rep)
Rich cohort study data[c]		
Mother's and partner's parenting scores at 6, 18, 24 and 38 months		Depression score at age 10
Mother/partner reads to child		IQ at age 8 (WISC scale)
SDQ score at age 6 (reported by mother) and age 7 (reported by teacher)		
School characteristics[d]		
School size	Percentage eligible for FSM	School type
Average class size	Percentage non-white	Percentage EAL
Duration head teacher in post		

[a]We use administrative data from the National Pupil Database (see http://nationalpupildatabase.wikispaces.com/ for more details on this data-set.).
[b]These (or similar) measures are available in other longitudinal surveys in the UK, such as the Longitudinal Study of Young People in England (see http://www.esds.ac.uk/longitudinal/access/lsype/L5545.asp for more details on this data-set.), although our measures are taken from ALSPAC.
[c]We use variables from ALSPAC.
[d]School-level data are mainly available from administrative sources (National Pupil Database [NPD]) with information on the duration that a head teacher has been in post taken from ALSPAC.
Note: SDQ, strengths and difficulties questionnaire; WISC, Weschler Intelligence Scale for Children; EAL, English as an additional language.

Table 2. Descriptive statistics for explanatory variables, comparing the analysis sample with the population of English pupils.

	ALSPAC analysis sample	NPD analysis sample
Administrative data		
Achieved expected level in English at age 11	80.7%	79.0%
Achieved expected level in Maths at age 11	78.6%	76.1%
Achieved expected level in Science at age 11	91.3%	89.4%
Eligible for FSM	10.3%	16.4%
Non-statemented SEN	16.9%	20.2%
White British ethnic origin	95.6%	83.7%
Typical longitudinal survey data		
Mother has at least O-level qualifications	67.2%	
Mother has a degree	8.6%	
Partner has at least O-level qualifications	63.4%	
Partner has a degree	12.7%	
Child has ever lived in a single parent family	9.3%	
Child has ever lived in social housing	20.0%	
Child was breastfed	70.0%	
Rich cohort study data		
Mother frequently reads to child	62.6%	
Partner frequently reads to child	28.2%	
School characteristics		
Attends a community school	66.3%	70.0%
Average school size	294 pupils	220 pupils
Observations	5417	1,635,573

progress between ages 7 and 11 initially using the rich ALSPAC data and then using the sparse administrative data.

We start by considering the ALSPAC data. As we control for more and more pupil-level factors, we demonstrate how the effect of SEN status varies as we account for differences in intake between the schools and, it is hoped, for the effects of how the pupils select their schools. Note that, in this case, 'pupil-level' selection involves decisions by parents about the schools they wish to send their children to, based on the family's circumstances and the child's specific needs.

The simplest model, M1, includes only age 7 test score and our treatment indicator, SEN status; model M2 includes further variables from a typically sparse administrative data source, such as gender, ethnicity and eligibility for free school meals; model M3 additionally includes measures from a typical longitudinal survey, such as richer measures of parents' socio-economic status and family composition; and finally, rich data from ALSPAC – such as IQ and measures of mother's and father's parenting skills – are included in M4. If we find differences between the RE and FE estimates, then it would suggest that unmeasured but important school-level influences on progress are correlated with SEN status, and the RE assumption that u_j is uncorrelated with SEN would be called into question.

To investigate the impact of school-level selection on the analysis and the validity of the RE assumption, we further include school-level variables in RE model M5. If the estimates of the impact of SEN status on pupil progress under the FE model M4 and RE model M5 are substantially different, then this indicates failure of the RE assumption. Moreover, as a secondary question, we use random coefficients models to examine whether there is heterogeneity in the effect of SEN status on progress between ages 7 and 11.

We then contrast the rich model above with a simply specified model using only the available administrative data. Specifically we estimate model M1, including only age 7 test score and SEN status, and then model M2, which includes the limited set of covariates set out in the top panel of Table 1.

4.4 *Effects of SEN status on progress*

On average, 16.9% of pupils in our ALSPAC sample are recorded as having non-statemented SEN, but there is variation across schools, with a quarter of schools having fewer than 15% of pupils labelled as having SEN and a further quarter with more than 24% of pupils labelled as having SEN. Table 3 gives a large negative effect of non-statemented SEN status on progress, regardless of whether school-level terms are treated as fixed or random. Pupils labelled as having non-statemented SEN score around 0.3 standard deviations lower in the age 11 tests than pupils with similar levels of prior achievement who are not labelled (M1). When using these ALSPAC data, the effect of SEN remains substantial even after controlling for an array of child and family measures (M2–M4).

In terms of the choice between RE and FE models, the two sets of estimates are very similar, with a relative difference of no more than 2.2% that falls to 0.6% with the inclusion of school-level variables in M5. This suggests that the variation in SEN across schools is driven by factors not associated with academic progress once pupil-level factors have been controlled for. For each model, the Hausman test fails to reject the null hypothesis that the fixed and random effect estimates are estimating the same parameter, with the *p*-value increasing as more variables are added to the model. We can therefore conclude that the RE assumption holds in this case, possibly because the rich survey data have allowed us to adjust for selection.

The similarity of the FE and RE estimates together with the Hausman test results suggest that we can adopt the RE approach in this case. We can thus consider extending the simple random intercept models we have been using up until this point. In

Table 3. Estimated effects of SEN status on progress between ages 7 and 11 for various FE and RE specifications using ALSPAC data.

Model	FE		RE			% diff.[b]	Hausman *p*-value[c]
	$\hat{\beta}_{FE}$	(se)	$\hat{\beta}_{RE}$	(se)	$\hat{\rho}_{RE}$[a]		
M1. KS1 average point score only	−0.335	(0.025)	−0.330	(0.025)	0.157	1.5	0.113
M2: M1 + administrative data	−0.347	(0.025)	−0.342	(0.025)	0.154	1.4	0.213
M3. M2 + typical survey data	−0.355	(0.025)	−0.351	(0.024)	0.137	1.7	0.384
M4: M3 + rich cohort data	−0.321	(0.024)	−0.316	(0.024)	0.139	2.2	0.358
M5: M4 + school-level data	−0.321	(0.024)	−0.320	(0.024)	0.117	0.6	0.794
Number of pupils	5417						
Number of schools	200						

[a] $\hat{\rho}_{RE}$ is the intra-school correlation estimated from the RE model as $\hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \hat{\sigma}_e^2)$.
[b] The relative difference between the FE and RE estimates is calculated as $100 \times (\hat{\beta}_{RE} - \hat{\beta}_{FE})/\hat{\beta}_{FE}$.
[c] *p*-value for Hausman test statistic for SEN effect as given by Equation (4).

Table 4. Estimated variance components for models with (i) random intercepts only (M5 of Table 3) and (ii) random coefficient for SEN status.

	Random intercept		Random coefficient for SEN	
	Est	(se)	Est	(se)
School level				
Intercept variance (σ^2_{u0})	0.037	(0.005)	0.029	(0.005)
SEN effect variance (σ^2_{u1})	–	–	0.149	(0.028)
Intercept-SEN covariance (σ_{u01})	–	–	0.017	(0.009)
Pupil level				
σ^2_e	0.282	(0.006)	0.266	(0.005)
-log-likelihood	4388		4318	

particular, we can consider a random coefficients model for SEN status in order to investigate heterogeneity in the effect of SEN status across schools. Table 4 gives estimates of the variance components from M5 and its extension to allow a random coefficient for SEN. There is strong evidence that the effect of SEN varies across schools (LR statistic = 140, 2 df, $p < 0.0001$). Denoting the school-level random intercept by u_{0j} and the random coefficient of SEN by u_{1j}, with variances σ^2_{u0} and σ^2_{u1} and covariance σ_{u01}, it can be shown that the between-school variance is

$$\text{var}\left(u_{0j} + u_{1j}\text{SEN}_{ij}\right) = \sigma^2_{u0} + 2\sigma_{u01}\text{SEN}_{ij} + \sigma^2_{u1}\text{SEN}^2_{ij}, \tag{9}$$

which, because SEN_{ij} is binary, simplifies to σ^2_{u0} for non-SEN pupils and $\sigma^2_{u0} + 2\sigma_{u01} + \sigma^2_{u1}$ for SEN pupils. Substituting the estimates of the school-level variance components from Table 4, we obtain a between-school variance of 0.029 for non-SEN and 0.212 for SEN pupils, which implies that the school effects are substantially larger for SEN than for non-SEN pupils. The heterogeneity indicates that there are substantial differences between schools in the types of SEN policies used, the effectiveness with which these interventions are implemented and/or the policies by which pupils are selected into the SEN treatment; moreover, if the regression assumption has failed then the sources of heterogeneity will also include between-school differences in pupil intake.

The similarity of the estimated effect of SEN between models M4 and M5 suggests that the dominant effects of selection appear to be working at the pupil level. However, although we can be fairly sure about the RE assumption, it remains to be established whether the negative effect of SEN is a policy-relevant estimate. For this interpretation to hold, we must assume that the regression assumption holds and that all the individual-level variables strongly associated with selection have been included in the model. Of course, this is a bold assumption but one that cannot be tested using the available data. To be more certain of our results, we might use a quasi-experimental approach based on instrumental variables or regression discontinuity designs (e.g. Shadish, Cook, and Campbell 2002), but it is not possible to use either approach for this analysis. We can, however, compare our results with those obtained using propensity scores, though we are mindful that propensity score matching also relies on the assumption that selection is on the basis of observable characteristics. On the other hand, propensity scores do not rely on any regression model for the outcome and can be used to directly compare treated and untreated pupils who are virtually the same (in terms of their

Table 5. Estimated effects of SEN status on progress between ages 7 and 11 for various FE and RE specifications using NPD data.

Model	FE		RE				Hausman
	$\hat{\beta}_{FE}$	(se)	$\hat{\beta}_{RE}$	(se)	$\hat{\rho}_{RE}$[a]	% diff.[b]	p-value[c]
M1. KS1 average point score only	−0.456	(0.001)	−0.457	(0.001)	0.138	0.2%	0.000
M2: M1 + administrative data	−0.465	(0.001)	−0.466	(0.001)	0.121	0.2%	0.000
Number of pupils	1,635,573						
Number of schools	16,875						

[a]$\hat{\rho}_{RE}$ is the intra-school correlation estimated from the RE model as $\hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \hat{\sigma}_e^2)$.
[b]The relative difference between the FE and RE estimates is calculated as $100 \times (\hat{\beta}_{RE} - \hat{\beta}_{FE})/\hat{\beta}_{FE}$.
[c]p-value for Hausman test statistic for SEN effect as given by Equation (4).

observed characteristics) except for their SEN status. Crawford and Vignoles (2010) used a propensity score matching approach that attaches the greatest weight to matches between SEN pupils and non-SEN pupils who, while not labelled SEN by the school system, are judged to have special education needs by their teachers. Their more detailed analysis supports the conclusions drawn from this paper, namely, pupils labelled as having SEN make significantly less progress than otherwise observationally identical pupils without such labels.

We now turn to the second example based on the National Pupil Database in which we use administrative data only. The proportion of pupils labelled as having SEN is slightly larger for the whole cohort than it was for the ALSPAC sample: 20.2% compared to 16.9%. There is also slightly more variation across schools, with a quarter of schools in the whole NPD having fewer than 13% of pupils labelled as having SEN and a further quarter having more than 26%; the equivalent figures for the ALSPAC sample were 15% and 24%, respectively.

Table 5 contains the results for both model M1, which controls only for SEN status and Key Stage 1 test scores at age 7, and model M2, which additionally controls for a limited set of background characteristics such as eligibility for free school meals. We can see that the Hausman test rejects at the 1% level of significance the null hypothesis that the FE and RE models are estimating the same parameters. Compared to the first example, the available data are much sparser (in the sense of having only limited information available on each pupil) and so it is more likely that the model is unable to fully account for the selection of pupils into schools. Hence, it would appear unlikely that the RE assumption holds. Substantively, however, it is clear that the estimated relationship between SEN status and Key Stage 2 scores is qualitatively similar regardless of whether we use fixed or random school effects. In this case, the rejection of the Hausman test is largely driven by the large sample sizes and the precision of the model.

It is interesting to note, however, that the coefficient on SEN status is about one-third larger for the full NPD cohort than was the case for the ALSPAC sample: pupils labelled as having SEN in the full NPD cohort score around 45% of a standard deviation less than otherwise similar pupils without a SEN label, while they scored around 35% of a standard deviation lower in the ALSPAC sample. This may indicate that schools in the Avon area of England are more successful at 'treating' SEN pupils and hence that these pupils make relatively more progress than SEN pupils in other areas of England. Alternatively, the differences might arise from differential sample selection. Certainly sample characteristics vary across the two sets of data in terms of the age range of students, ethnicity and other factors.

Table 6. Estimated effects of FSM eligibility on progress between ages 7 and 11 for various FE and RE specifications using NPD data.

Model	FE		RE			% diff.[b]	Hausman *p*-value[c]
	$\hat{\beta}_{FE}$	(se)	$\hat{\beta}_{RE}$	(se)	$\hat{\rho}_{RE}$[a]		
M1. KS1 average point score only	−0.158	(0.001)	−0.163	(0.001)	0.123	3.2%	0.000
M2: M1 + administrative data	−0.134	(0.001)	−0.140	(0.001)	0.121	4.5%	0.000
Number of pupils	1,635,573						
Number of schools	16,875						

[a]$\hat{\rho}_{RE}$ is the intra-school correlation estimated from the RE model as $\hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \hat{\sigma}_e^2)$.
[b]The relative difference between the FE and RE estimates is calculated as $100 \times (\hat{\beta}_{RE} - \hat{\beta}_{FE})/\hat{\beta}_{FE}$.
[c]*p*-value for Hausman test statistic for SEN effect as given by Equation (4).

Crucially, for some other variables in the model, there is a more substantive difference between the coefficients from the RE and FE models. For example, Table 6 gives the relationship between eligibility for free school meals and pupil achievement (based on the same models given in Table 5). For completeness, M1 in this table is a model including only the FSM variable and the Key Stage 1 test scores and M2 is exactly the same model as M2 given in Table 5. Note that the FSM variable is included in the model as an indicator of socio-economic disadvantage, based on the large literature suggesting a strong relationship between socio-economic disadvantage and lower pupil achievement. In both M1 and M2, the magnitude of the coefficient on FSM eligibility is appreciably smaller with the preferred FE model. Specifically, the magnitude of the coefficient on the FSM variable is 4.5% smaller when using the more appropriate FE model (M2), suggesting a somewhat weaker relationship between socio-economic disadvantage and pupil achievement than might be implied by a RE model. This is consistent with selection into schools on the basis of unobservable characteristics correlated with both FSM and pupil achievement. This example illustrates that choice of model can be important from a substantive perspective.

5. Discussion

The substantive findings from this paper suggest that children who are identified as having SEN (but who do not have a 'statement') achieve around 35–45% of a standard deviation less in Key Stage 2 tests than otherwise similar pupils who do not have a SEN label. We were unable to find positive evidence that being identified as having SEN is associated with doing better at school academically (see Crawford and Vignoles 2010 for further investigation of this issue).

The primary aim of this paper, however, has been to highlight the key issues that should be considered when deciding whether to treat the school-level terms in a hierarchical regression as FE or RE. We hope to have clarified for social and education researchers less familiar with these issues why economists tend to prefer FE models. Similarly, we hope to have clarified that FE models are not a panacea: adjustment for school-level variables can mitigate against the failure of RE assumption and FE are not robust to failure of the regression (or exogeneity) assumption. Further, RE models have the advantage of greater efficiency, which may be important particularly in cases where some schools have small numbers of pupils and estimates of the

regression coefficients and school-level residuals may be imprecise. RE models also enable researchers to identify the effects of higher level characteristics, something which is clearly appealing for education researchers. While these methodological points are certainly not new, we hope that this paper encourages researchers from all disciplines not to be constrained by dogma in their choice of methods.

Two issues have been of concern throughout.

(1) The presence of correlation between unobserved individual characteristics (that are correlated with the outcome) and the treatment indicator; both FE and RE models assume that this correlation is zero, which we have referred to as the regression assumption.
(2) The presence of correlation between unobserved higher level characteristics (that are correlated with the outcome) and the treatment indicator; RE models (but not FE models) assume that this correlation is zero, which we have referred to as the RE assumption.

Both points are related to the non-random selection of individuals into higher level units. It is paramount that researchers account for both types of selection in their models. The choice between FE and RE models hinges on the second of these concerns. The FE approach will be preferable in scenarios where selection is insufficiently understood or the data have a limited range of variables with which selection can be adjusted: its robustness to the RE assumption is attractive and educational researchers should consider using a FE model in such scenarios, even if only to assess the robustness of estimates from an equivalent RE model. Indeed in our second example, using sparser administrative data, the Hausman rejected the RE assumption, and while this did not make a substantive difference to the coefficient on our SEN variable, it did make a substantive difference to the coefficients on other variables such as eligibility for FSM. However, when the available data on higher level units are rich, RE models can be built that adjust for higher level selection and the estimates checked against FE models for robustness to failure of the RE assumption; if robustness is indicated, then the estimates from RE models can be reliably reported and additional analyses of shrunken estimates of school residuals and random coefficients can be performed.

We also believe that it is important to take a pragmatic view of what can reasonably be achieved by analysing data from observational studies, whichever approach is used. Strictly, causal inferences require randomised interventions or, failing that, quasi-randomised experimental designs such as those based on instrumental variables and regression discontinuity. However, a realistic aim of analyses based on observational studies is inference that is policy-relevant; that is, estimates that do not lead to misleading policy recommendations. Even this limited aim requires the availability of rich data, preferably aided by background knowledge of the selection process.

We therefore conclude that since an assumption of selection on the basis of observables is often problematic, it is essential that to obtain policy-relevant estimates, researchers check the robustness of their results by comparing FE and RE models. Given the ease with which such models can be estimated there really is no excuse for researchers to continue to use one approach without checking whether their estimates are affected by their choice of model. Even if researchers do check RE estimates against FE estimates, no conclusion is infallible because other assumptions may be violated, but it is crucial, we would argue, that the assumptions under which policy

recommendations are made should be highlighted and subjected to critical scrutiny. It is also particularly important that these assumptions should form the basis for a common 'language' for criticism in areas of multidisciplinary research like education.

Acknowledgements

The authors would like to thank Rebecca Allen, Simon Burgess, seminar participants at the University of Bristol and the Institute of Education and two anonymous referees for their comments and advice. All errors remain the responsibility of the authors.

Funding

The authors gratefully acknowledge funding from the Economic & Social Research Council [grant number RES-060-23-0011].

Notes

1. Iterative or feasible generalised least squares estimators of random effects models are normally distributed in large samples, whether or not the residuals/errors are normally distributed. However, note that most maximum likelihood and Markov chain Monte Carlo estimators do assume residual normality, but linearity means that non-normality affects only estimator efficiency rather than bias.
2. Technically, this assumption is that the population average of the u_j given the covariates is zero, but it almost always corresponds to assuming that the correlation between the covariates and the school-level residuals is zero.
3. Strictly speaking, $\boldsymbol{w_j}$ and $\boldsymbol{z_j}$ are both mean-centred variables to ensure the population average of u_j is zero.
4. In the scalar case, the school-mean model (8) requires only that $w_j = g_0 + g_1\bar{x}_j + v_j$ with $g_1 \neq 0$, where v_j is a residual independent of x_{ij}. However, if, for example, the true relationship is $w_j = \gamma_0 + \gamma_1\bar{x}_j + \gamma_2\bar{x}_j^2 + \omega_j$, then $u_j^* = \{\gamma_0 - g_0 + (\gamma_1 - g_1)\bar{x}_j + \gamma_2\bar{x}_j^2 + \omega_j\}\delta + z_j\lambda$, which confounds the effect of non-linearity and residual ω_j with the desired $z_j\lambda$.
5. The relationship between SEN status and pupil progress is considered in more detail in Crawford and Vignoles (2010).
6. SEN status can change over time and this variation has been used by other authors (e.g. Hanushek, Kain, and Rivkin 2002; Meschi and Vignoles 2010) to identify the effect of SEN status on educational attainment using individual FE models. In our example, SEN status is measured at age 10, but the estimates vary little if we use SEN status measured at age 7 instead.
7. See http://www.bristol.ac.uk/alspac/sci-com/ for more details on the ALSPAC data resource.
8. See http://www.sdqinfo.com/b1.html for more details.

References

Aitkin, M., and N. Longford. 1986. "Statistical Modelling in School Effectiveness Studies." *Journal of the Royal Statistical Society* 149 (part 1): 1–43.

Blatchford, P., P. Bassett, P. Brown, C. Martin, A. Russell, and R. Webster. 2011. "The Impact of Support Staff on Pupils' 'Positive Approaches to Learning' and Their Academic Progress." *British Educational Research Journal* 37 (3): 443–464.

Blundell, R., L. Dearden, and B. Sianesi. 2005. "Evaluating the Effect of Education on Earnings: Models, Methods and Results from the National Child Development Survey." *Journal of the Royal Statistical Society*, Series A, 168 (3): 473–512.

Burgess, S., E. Greaves, A. Vignoles, and D. Wilson. 2011. "Parental Choice of Primary School in England: What Types of School Do Different Types of Family Really Have Available to Them?" *Policy Studies* 32 (5): 531–547.

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.
Crawford, C., and A. Vignoles. 2010. *An Analysis of the Educational Progress of Children Classified as Having Special Educational Needs*. DoQSS Working Paper No. 10–19, Department of Quantitative Social Science, Institute of Education, UK.
Ebbes, P., U. Bockenholt, and M. Wedel. 2004. "Regressor and Random-Effects Dependencies in Multilevel Models." *Statistica Neerlandica* 58 (2): 161–178.
Efron, B., and C. Morris. 1973. "Stein's Estimation Rule and Its Competitors – An Empirical Bayes Approach." *Journal of the American Statistical Association* 68 (1): 117–130.
Fielding, A. 2004. "The Role of the Hausman Test and Whether Higher Level Effects Should Be Treated as Random or Fixed." *Multilevel Modelling Newsletter* 16 (2). http://www.bristol.ac.uk/cmm/learning/support/new16–2.pdf
Gelman, A., and J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.
Goldstein, H. 1997. "Methods in School Effectiveness Research." *School Effectiveness and School Improvement* 8 (4): 369–395.
Goldstein, H. 2010. *Multilevel Statistical Models*. 4th ed. London: Wiley.
Grilli, L., and C. Rampichini. 2011. "The Role of Sample Cluster Means in Multilevel Models: A View on Endogeneity and Measurement Error Issues." *Methodology* 7 (4): 121–133.
Hanushek, E., J. Kain, and S. Rivkin. 2002. "Inferring Program Effects for Special Populations: Does Special Education Raise Achievement for Students with Disabilities?" *Review of Economics and Statistics* 84 (4): 584–599.
Hong, G., and S. W. Raudenbush. 2006. "Evaluating Kindergarten Retention Policy: A Case Study of Multilevel Inference for Multi-level Observational Data." *Journal of the American Statistical Association* 101 (475): 901–910.
Lamb, B. 2010. *The Lamb Inquiry: Special Educational Needs and Parental Confidence*. Nottingham: DCSF Publications. http://www.dcsf.gov.uk/lambinquiry/downloads/8553-lamb-inquiry.pdf
Meschi, E., and A. Vignoles. 2010. *An Investigation of Pupils with Speech, Language and Communication Needs (SLCN)*. DCSF Research Report.
Mundlak, Y. 1978. "On the Pooling of Time Series and Cross Section Data." *Econometrica* 46 (1): 69–85.
Nuttall, D. L., H. Goldstein, R. Prosser, and J. Rasbash. 1989. "Differential School Effectiveness." *International Journal of Educational Research* 13 (7): 769–776.
Ofsted. 2010. *The Special Educational Needs and Disability Review*. HMI: 090221. http://www.ofsted.gov.uk/Ofsted-home/Publications-and-research/Browse-all-by/Documents-by-type/Thematic-reports/The-special-educational-needs-and-disability-review
Sammons, P., D. Nuttal, and P. Cuttance. 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.
Shadish, W. R., T. D. Cook, and D. T. Campbell. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.
Skrondal, A., and S. Rabe-Hesketh. 2004. *Generalized Latent Variable Modelling: Multilevel, Longitudinal, and Structural Equation Models*, pp. 52–53. Boca Raton, FL: Chapman & Hall/CRC.
Snijders, T .A. B., and J. Berkhof. 2008. "Diagnostic Checks for Multilevel Models, Chapter 3." In *Handbook of Multilevel Analysis*, edited by J. De Leeuw and E. Meijer, 141–175. Thousand Oaks, CA: Sage.
Snijders, T. A. B., and R. J. Bosker. 2011. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling*. London: Sage.
Webster, R., and P. Blatchford. 2013. *The Making a Statement Project. Final Report. A Study of the Teaching and Support Experienced by Pupils with a Statement of Special Educational Needs in Mainstream Primary Schools*. Department of Psychology and Human Development, Institute of Education, UK. http://www.schoolsupportstaff.net/mastreport.pdf
Woodhouse, G., M. Yang, H. Goldstein, and J. Rasbash. 1996. "Adjusting for Measurement-Error in Multilevel Analysis." *Journal of the Royal Statistical Societ*o, Series A, 159 (2): 201–212.
Wooldridge, J. M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.