Fixed and Random effects: making an informed choice

Andrew Bell
The University of Sheffield

Malcolm Fairbrother
University of Bristol

Kelvyn Jones
University of Bristol

Fixed and Random effects: making an informed choice

Andrew Bell[1], Malcolm Fairbrother[2], Kelvyn Jones[2]

[1] Sheffield Methods Institute, University of Sheffield

[2] School of Geographical Sciences, University of Bristol

Last updated: 28th March 2016

Email: andrew.j.d.bell@sheffield.ac.uk

Abstract

This paper considers the modelling choices available to researchers using multilevel data, including longitudinal data of various types. Specifically, we consider fixed effects (FE) and random effects (RE) models, including the within-between RE model, often misleadingly termed the 'hybrid' model. We argue that the latter is unambiguously a RE model, and that it is the most general of the three models, and as such a sensible starting point, given its flexibility to incorporate the positive aspects of both FE and RE models, and its ability to allow extensions (such as random slopes) that are often important. We present simulations that reveal the extent to which these models cope with mis-specification, finding that failing to include random slopes (e.g. in a FE or standard RE model) can lead to anti-conservative standard errors, and that mis-specifying non-Normal random effects as Normally distributed can introduce some small biases to variance and random effect estimates, but not fixed-part parameter estimates. We conclude with advice for applied researchers, and present a glossary, that gives clear definitions to terms that are confusing or have more than one meaning. Overall, we hope the paper gives practical advice to researchers in many different social science disciplines and beyond, looking to understand and use multilevel and longitudinal data.

1 Introduction

Analyses of longitudinal data and other data with multiple levels can employ a variety of different methods given the structures present in such data. However, in our view there is significant confusion regarding these various methods, including what they each can show. This paper aims to present a comprehensive review of the difference between two key methods: fixed effects (FE) models and random effects (RE) models. We have argued before (Bell and Jones 2015) that in many if not most research scenarios, RE provides everything that FE provides and more, making it the preferred method for many practitioners (see also Shor et al. 2007; Western 1998). However, this view is at odds with widespread impressions that FE is the gold standard in such contexts.

This paper serves in part as a commentary on a recent contribution which we regard as valuable in many ways, but also somewhat misleading (Vaisey and Miles 2014). It also addresses broader issues arising from prevalent misunderstandings about FE and RE models: how they can be used, what they do, what they assume, and even the fundamental question of what they are. The literature is rife with confusing terminology (such as the phrase 'random effects' itself—see for example Gelman 2005), and different disciplines exhibit different approaches to the same important methodological questions.

Much of what we will argue here is not new; indeed much of it has been a part of the published research literature for close to 40 years. This paper's key contribution is to synthesize this work, which has been conducted in a range of disciplines (sociology, political science, economics, geography, linguistics, statistics, and others), and to provide a more complete picture of issues that have only been partially addressed in individual disciplines. We also help answer some previously unanswered questions with a new simulation study.

The paper begins by outlining the data structure that both FE and RE aim to capture: clustering or dependence in the data, and differing relationships within- and between-clusters. We then present our favoured model: a RE model that allows for distinct within and between 'effects'[1], which we abbreviate "REWB". We show how the more commonly used FE, RE and pooled OLS models can be understood as constrained versions of this model; indeed REWB is our favoured model because of its encompassing nature. We then go on to consider some important extensions to this model that cannot be implemented under a FE or OLS framework: 'random slopes' models that allow for the associations between variables to vary across higher-level units, the addition of further spatial and temporal levels of analysis,

[1] The use of the term 'effect' in the phrase 'within effect', 'between effect' and 'contextual effect' should not imply that these should necessarily be interpreted as causal. This caution applies to the phrases 'random effects' and 'fixed effects' as well.

and the explicit modelling of complex level 1 heteroscedasticity. We show that implementing these extensions can often be of paramount importance and can make results both more nuanced and more accurate, in ways that cannot be considered under the FE framework. The paper then considers models with a non-continuous dependent variable, and some of the different challenges that such models present, before considering the assumptions made by the RE model and the extent to which it matters when those assumptions are violated. The article concludes with some practical guidelines designed to help researchers in deciding what model they should use and how.

At the end of this article there is a glossary, which we hope will serve to clarify the multiple meanings attached to some key terms. This is a list of terms that either (a) have more than one meaning, or (b) that have meanings that are commonly confused, and as such is not meant to be exhaustive.

2 Within, between and contextual effects: conceptualising the fixed part of the model

Social science datasets often have complex structures, and these structures can be highly relevant to the research question at hand. Often, observations (at level 1) are clustered into groups of some kind (at level 2). Such two-level data structures are the main focus of this paper, though data are sometimes grouped at further levels, yielding three (or more) levels. Some of the most common multilevel structures are outlined in table 1.[2] In broad terms, these can be categorised into two types: cross-sectional data, where individuals are nested within a geographical or social context (e.g. individuals within schools or countries), and longitudinal data, where individuals or social units are measured on a number of occasions. In the latter context, this means occasions are nested within the individual or unit. In all

[2] It is also possible to fit a multivariate model in which multiple dependent variables form the lowest level; for an example see Deeming and Jones (2015).

cases these structures represent real and interesting societal configurations; they are not simply a technicality or consequence of survey methodology.

Structures are important in part because variables can be related at more than one level in a hierarchical data structure, and the relationships at different levels are not necessarily equivalent. Taking a cross-sectional example: some social attitude (Y) may be related to an individual's income X (at level 1) very differently than to the average income in their neighbourhood or country (level 2). As a panel example: an individual might be affected by having what is, for them, an unusually high income (level 1) in a different way to the effect of being an individual who has a high income generally (level 2). Thus, we can have "within" effects that occur at level 1 within neighbourhoods (cross-sectional) or within individuals (panel), and "between" or "contextual" effects that occur at level 2 between neighbourhoods (cross-sectional) or individuals (panel) (Howard 2015).

Sometimes it is the case that within effects are of greatest interest. With panel data, within effects can represent the effect of changing the value of an independent variable over time. When examining the effect of some policy intervention, it can be particularly illuminating to investigate the relationship between the implementation of some policy intervention and a change in the dependent variable. Many studies have argued for focusing on the longitudinal relationships in some dataset because unobserved, time-invariant differences among the units measured repeatedly over time are controlled for (Allison 1994; Halaby 2004).

Yet in many research contexts there may well also be unobserved changes in the circumstances of such units; there is no guarantee that within relationships are unaffected by omitted variable bias. Between effects can be equally illuminating, despite being by definition non-changing—as evidenced by the many published studies that rely exclusively on cross-sectional data. Social science is concerned with understanding the world as it exists, not just dynamic changes in it. Thus with a panel dataset for example, it will often be worth modelling associations at the higher level, in order to understand the ways in which individuals differ – not just the ways in which they change (see, for example, Bell and Jones 2015; Subramanian et al. 2009). We take it as axiomatic that we need both micro and macro associations to understand the whole of 'what is going on'.

Table 1: Some hierarchical structures of data that are common in social science. For more elaboration of hierarchical and non-hierarchical structures, see (Rasbash 2008).

Broad category	Data type	Level 1	Level 2	Level 3
Cross-sectional	Clustered survey data (Maimon and Kuhl 2008)	Individuals	Neighborhoods	-
Cross-sectional	Cross-national survey data (Ruiter and van Tubergen 2009)	Individuals	Countries	-
Cross-sectional	Surveys with multiple items (Sampson, Raudenbush, and Earls 1997)	Items	Individuals	-
Panel	Country time-series cross-sectional data (Beck and Katz 1995; Walder, Isaacson, and Lu 2015; Western 1998)	Occasions	Countries	-
Panel	Individual panel data (Lauen and Gaddis 2013)	Occasions	Individuals	-
Panel at level 1, cross-sectional at level 2	Panel data on individuals who are clustered (Kloosterman et al. 2010)	Occasions	Individuals	Schools
Cross-sectional at level 1, Panel at level 2	Comparative longitudinal survey data (Fairbrother 2014; Schmidt-Catran and Spies 2016)	Individuals	Country-Years/ Region-Years	Countries/ Regions

2.1 The most general: within-between RE and Mundlak models

We now outline some statistical models that aim to represent these processes. Taking a panel data example, where individuals i are measured on multiple occasions t, we can conceive of the following model—the most general of the models that we present in this section of the paper, in that it is able to model both within- and between-individual effects concurrently:

$$y_{it} = \beta_0 + \beta_1(x_{it} - \bar{x}_i) + \beta_2\bar{x}_i + \beta_4 z_i + (\upsilon_i + \epsilon_{it}). \tag{1}$$

Here, y_{it} is the dependent variable, x_{it} is a time-varying (level 1) independent variable, and z_i is a time-invariant (level 2) independent variable. β_1 represents the within effect of x_{it}, whilst β_2 represents the between effect[3] of x_{it}. The β_4 parameter represents the effect of time-invariant variable z_i, and is therefore in itself a between effect (since level 2 variables cannot have within effects since there is no variation within higher level units). u_i are the model's random effects, which are assumed to be Normally distributed (see section 3 for more on this). ϵ_{it} are the model's level 1 residuals, which are also assumed to be Normally distributed (we will discuss models for non-Gaussian outcomes, with different distributional assumptions, later).

An alternative parameterisation to equation 1 (with the same distributional assumptions) is the 'Mundlak' formulation (Mundlak 1978):

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_3\bar{x}_i + \beta_4 z_i + (\upsilon_i + \epsilon_{it}). \tag{2}$$

[3] Note that the variable $\bar{x}_i$ associated with β_2 could be calculated using only observations for which there is a full data record, though if more data exists this could be included in the calculation of $\bar{x}_i$, to improve the estimate of β_2. Alternatively, calculating $(x_{it} - \bar{x}_i)$ with only observations included in the model ensures β_1 is estimated using only within-unit variation. In practice, the difference between these modelling choices is usually negligible.

With x_{ij} included in its raw form rather than de-meaned form $(x_{it} - \bar{x}_i)$. Instead of the between effect β_2, we are estimating the "contextual effect" β_3. The key difference between these two, as spelled out by Raudenbush and Bryk (2002:140) is that the raw value of the time–varying predictor (x_{it}) is controlled for in the estimate of the contextual effect in equation 2, but not in the estimate of the between effect in equation 1. Thus if the research question at hand is "what is the effect of an individual moving from one level 2 unit to another", the contextual effect (β_3) is of more interest, since it holds the level 1 individual characteristics constant. In contrast, if we simply want to know "what is the difference between two level 2 units", the between effect (β_2) will provide an answer to that. In general, when using cross-sectional data, the contextual effect is of interest (since we can imagine level 1 individuals moving between level 2 units without altering their own characteristics), whereas with panel data (where level 1 units, occasions, are rather abstract and often of little interest in themselves) the between effect is generally more informative. Note however, that these models are equivalent, since $\beta_1 + \beta_3 = \beta_2$; each model conveys the same information and will fit the data equally well and we can obtain one from the other with some simple arithmetic[4].

Both the Mundlak model and the within-between random effects (REWB) models (equations 2 and 1 respectively) are easy to fit in all major software packages (e.g. R, Stata, SAS, as well as more specialist software like HLM and MLwiN). They are simply random effects models with the mean of x_{ij} included as an additional explanatory variable (Howard 2015).

2.2 Constraining the within-between RE model: fixed effects, random effects and OLS.

Having established our 'encompassing' model in its two alternative forms (Mundlak, and within-between), we now present three models that are currently more often used. In

[4] One potential advantage of the within-between model is that there will be zero correlation between $\bar{x}_i$ and $(x_{it} - \bar{x}_t)$, which can facilitate model convergence.

showing that each of these are constrained versions of equation 1 above, we question why one would not choose the more general and potentially more informative and revealing specification.

2.2.1 Random effects without within and between separation

One commonly used model uses the random effects framework, but does not estimate separate relationships at each of the two levels:

$$y_{it} = \beta_0 + \beta_1^{RE} x_{it} + \beta_4^{RE} z_i + (\upsilon_i + \epsilon_{it}) \tag{3}$$

This approach effectively assumes that $\beta_1 = \beta_2$, or that $\beta_3 = 0$, in equations 1 and 2. Where this assumption is valid, this model is a good choice, and has benefits over the more general model. Specifically, the estimate of β_1^{RE} will be more efficient than the estimates of β_1 or β_2 in equation 1, because it can utilise variation at both the higher and lower level (e.g., Fairbrother 2014; Halaby 2004). However, when $\beta_1 \neq \beta_2$, the model will produce a weighted average of the two[5], which will have little substantive meaning (Raudenbush and Bryk 2002:138). Fortunately it is easy to test whether the assumption of equal within and between effects is true, by testing the equality of the coefficients in the REWB model), or the significance of the contextual effect in the Mundlak model (for example via a Wald test). If there is a significant difference (and not just that the between effect is significant different from zero) the terms should not be combined, and the encompassing within-between or Mundlak model should be used. This was done by Hanchane and Mostafa (2011) considering bias with this model for school (level 2) and student (level 1) performance. They

[5] Specifically, the estimate will be weighted as: $\beta_{ML} = \frac{w_W \beta_W + w_B \beta_B}{w_W + w_B}$, where w_W is precision of the within estimate, that is $w_W = \left(1 / (SE_{\beta W})^2 \right)$ and w_B is precision of the between estimate, $w_B = \left(1 / (SE_{\beta B})^2 \right)$.

Given the larger sample size (and therefore higher precision) of the within estimate, the model will often tend towards the within estimate. β_W and β_B are the within and between effects, respectively (estimated as β_1 and β_2 in equation 1).

found that in less selective school systems (Finland), there was little bias and a model like equation 3 was appropriate, whilst in more selective systems (UK and Germany) the more encompassing model of equation 2 was necessary to take account of schools' contexts and estimate student effects accurately.

This is, in fact, what is effectively done by the oft-used 'Hausman test' (Hausman 1978). Although often (mis)used as a test of whether fixed or random should be used, it is really a test of whether there is a contextual effect, or whether the between and within effects are different. This equates in the panel case to whether the changing within effect (e.g. being unusually well paid, such as after receiving a pay raise) is different from the cross-sectional effect (being well paid on average, over the course of the period of observation). Even when within and between effects are slightly different, it may be that the bias in the estimated effect is a price deemed worth paying for the gains in efficiency, depending on the research question at hand (Clark and Linzer 2015). Either way, it is important to test whether the multilevel model in its commonly applied form of equation 3 is an uninterpretable blend of two different processes.

2.2.2 Fixed effects model

Depending on the field, perhaps the most commonly used and recommended method of dealing with differing within and between effects as outlined above is 'fixed effects' modelling. This approach is equivalent to that represented in equations 1 and 2, except that u_j are specified as fixed effects: i.e. dummy variables are included for each higher level entity (less a reference category) and the v_i are not treated as draws from any kind of distribution. The result is that between effects (associations at the higher level) cannot be estimated, and the model can be reduced to:

$$y_{it} = \beta_1(x_{it} - \bar{x}_i) + (v_i + \epsilon_{it}). \tag{4}$$

Or reduced even further to:

$$(y_{it} - \bar{y}_i) = \beta_1(x_{it} - \bar{x}_i) + (\epsilon_{it}).$$

(5)

Thus, the model is able to provide an estimate of the within effect β_1, which is not biased by different between effects. This is of course what is achieved by the REWB model and the Mundlak model. However, such specifications can tell us nothing about the higher level entities in the model. This means that many questions cannot be answered by FE, and it can only ever present an incomplete picture of the substantive phenomenon represented by the model. If a researcher has no substantive interest in the between effects, their exclusion is perhaps unimportant, though even in such a case, for reasons discussed below, we think there are still reasons to disfavour the FE approach as the one and only valid approach. To be clear the REWB and Mundlak will give exactly the same results for the within effect (coefficient and standard error) as the FE model, but retains the between effect which can be informative.

2.2.3 Single level OLS regression

An even simpler option is to ignore the structure of the model entirely:

$$y_{it} = \beta_0 + \beta_1^{OLS} x_{it} + \beta_4^{OLS} z_i + (\epsilon_{it})$$

(6)

Thus, we assume that all observations in the dataset are conditionally independent. This has two problems. First, as with the standard RE model, the estimate of β_1^{OLS} will be a potentially uninterpretable weighted average[6] of the within and between effects (if they are not equal). Furthermore, if there are differences between higher level entities, standard errors will be estimated as if all observations are independent, and so will be generally underestimated,

[6] This will actually be a different weighted average to that produced by RE: it is weighted by the proportion of the variance in x_{it} that exists at each level, so where the within-unit variance of x_{it} is negligible, the estimate will be close to that of the between effect, and vice versa. More formally, $\beta_{SL} = (1 - \rho_x)\beta_W + \rho_x\beta_B$, where ρ_x is the proportion of the variance in x_{it} occurring at the higher level.

especially for parameters associated with higher level variables, including between and contextual effects. Fortunately, the necessity of modelling the nested structure can readily be evaluated, by running the model both with and without the higher level random effects and testing which is the better fitting model by a likelihood ratio test (Snijders and Bosker 2012:97), AIC, or BIC.

2.3 Omitted variable bias in the within-between RE model

We hope the discussion above has convinced readers of the superiority of the REWB model, except perhaps when the within and between effects are approximately equal, in which case the standard RE model (without separated within and between effects) might be preferable for reasons of efficiency[7]. Even then, the REWB model should be considered first, or as an alternative, since the equality of the within and between coefficients should not be assumed. Except for the simplicity of the model, there is nothing that FE offers that the REWB does not.

All of the models we consider here are subject to a variety of biases, such as if there is selection bias (Delgado-Rodríguez and Llorca 2004), or the direction of causality assumed by the model is wrong (e.g. see Bell, Johnston, and Jones 2015). Most significantly for our purposes here is the possibility of omitted variable bias.

As with fixed effects models, the REWB model avoids bias on level 1 coefficients as a result of level 2 omitted variables, or put another way, correlation between level 1 variables included in the model and the level 2 random effects – such biases are absorbed into the between effect. This has been confirmed by simulations (Bell and Jones 2015; Fairbrother 2014). When using panel data with repeated measures on individuals, unchanging and/or unmeasured characteristics of an individual (such as intelligence, ability, etc.) will be

[7] This is not necessarily the case, however: if there are substantive reasons for suspecting that the processes driving the two effects are different then it makes sense to use SEs that treat the processes as separate. Moreover, it may be that subsequent elaboration of the model (addition of variables, etc.) would lead to within and between effects diverging – researchers are best served by being cautious about combining the two.

controlled out of the estimate of the within effect. However, excluded time-varying characteristics can still cause biases at level 1, just as they can in a fixed effects model. Similarly, those unmeasured level 2 characteristics can cause bias in the estimates of between effects and effects of other level 2 variables.

This is a problem if we wish to know the direct causal effect of a level 2 variable. However, this does not mean that those estimated relationships are worthless. Indeed often we are not looking for the direct, causal effect of a level 2 variable, but see these variables as proxies for a range of unmeasured social processes, which might include those omitted variables themselves. As an example, in a panel data structure when considering the relationship between ethnicity (an unchanging, level 2 variable) and a dependent variable, we would not interpret any association found to be the direct causal effect of any particular genes or skin pigmentation; rather we are interested in the effects of the myriad of unmeasured social and cultural factors that are related to ethnicity. If a direct genetic effect is what we are looking for, then our estimates are likely to be 'biased', but we hope most reasonable researchers would not interpret such coefficients in this way. As long as we interpret any coefficient estimates with these unmeasured variables in mind, and are aware that such reasoning is as much conceptual and theoretical as it is empirical, such coefficients can be of great value in helping us to understand patterns in the world through a model-based approach. Note that if we are, in fact, interested in a direct causal effect and are concerned by possible omitted variables, then instrumental variable techniques can be employed within the RE framework (for example, see Chatelain and Ralf 2010; Steele, Vignoles, and Jenkins 2007).

3 Fixed and random effects: Conceptualising the random part of the model

We have shown the different results that can be gleaned from fixed effects and random effects models, and the problems of bias that can arise from each of them. This section aims to clarify further the statistical and conceptual differences between RE and FE modelling frameworks.

The obvious statistical difference between the two models is in the way that that the higher level entities are treated. In a RE model (whether standard, REWB or Mundlak) clusters are treated as a random sample that is assumed to be from a Normal distribution, the variance of which is estimated:

$$\upsilon_i \sim N(0, \sigma_\upsilon^2). \tag{7}$$

In contrast, in a FE model, higher level entities are treated as completely independent:

$$\upsilon_i = \sum_{i=1}^{i} \beta_{0i} D_i \tag{8}$$

Here, D_i are dummy variables for each higher level entity i, each with a separately estimated coefficient β_{0i} in the fixed part of the model (less a reference category, or with the intercept suppressed). In both specifications, the level 1 variance is typically assumed to follow a Normal distribution:

$$\epsilon_{it} \sim N(0, \sigma_\epsilon^2) \tag{9}$$

In contrast, however, Vaisey and Miles (2014) argue that *the defining feature* of the RE model is an assumption that that model makes:

> "The only difference between RE and FE lies in the assumption they make about the relationship between *υ* [the unobserved time-constant fixed/random effects] and the observed predictors: *RE models assume that the observed predictors in the model are not correlated with υ while FE models allow them to be correlated"* (p. 4, italics in the original, underlining added).

Such views are also characteristic of mainstream econometrics:

> In modern econometric parlance, ‘‘random effect’’ is synonymous with zero correlation between the observed explanatory variables and the unobserved effect … the term ‘‘fixed effect’’ does not usually mean that c_i [υ_i in our notation] is being treated as nonrandom; rather, it means that one is allowing for arbitrary correlation between the unobserved effect c_i and the observed explanatory variables x_{it}. So, if c_i is called an ‘‘individual fixed effect’’ or a ‘‘firm fixed effect,’’ then, for practical purposes, this terminology means that c_i is allowed to be correlated with x_{it}. (Wooldridge 2002:252)

We do not doubt that this assumption is important (see section 2.3, above). But regardless of how well established this definition is, we hope that the discussion above shows why this definition is misleading and more confusing than it is helpful. This assumption is *not* the only difference between RE and FE models, and is far from being either model's defining feature. To us, a FE panel model treats person level unobservables as Fixed Effects (with dummy variables, one for each cluster, included in the model), whilst a RE model treats them as Random Effects drawn from a distribution. Any differences in assumptions between these models are incidental to the definition – they are related to what the model does, not what the model is.

The different distributional assumptions affect the extent to which information is considered to be exchangeable between higher level units: are units unrelated, or is the value of one higher level unit related to the values of the others? The FE framework effectively assumes that nothing can be known about each higher-level unit from any or all of the others—they are unrelated each exist completely independently. At the other extreme, a single-level model assumes there are no differences between the higher-level units, in a sense assuming that knowing one is sufficient to know them all. RE models strike an empirically identified balance between these two extremes, in assuming that higher-level units are distinct but of a kind—not completely unlike each other. The level-2 unit random effects in RE models come from a distribution, whereas the level-2 fixed effects in FE models do not, and the single-level model has no level-2 units. In practice, the random intercepts in RE models will correlate strongly with the fixed effects in FE models, but they will be drawn in towards their mean—with unreliably estimated outliers drawn in, or "shrunk", the most. The degree of this "shrinkage" (or exchangeability across level 2 units) is determined from the data, with more shrinkage if there are not many observations in a unit and/or the estimated variance of the higher level units, σ_{υ}^{2}, is small (see Spiegelhalter 2004).

Why does it matter that the random effects are drawn from a common distribution? In one sense, it doesn't matter a great deal. Using the REWB model will yield the same within-unit result as using an FE model, regardless of the nature of the given sample of clusters. Both methods will produce results that are not representative of the population if the sample is not representative of the population. Thus, the claim that FE conditions out the effects of higher level units is somewhat misleading. If the sample of units is non-random, and the effect of an independent variable of interest is heterogeneous across those units, both FE and REWB models will produce estimates that produce an average effect across the sample, not the population. Such heterogeneity can however be explicitly modelled within the RE framework as we show in section 4.1 below. Even when the sample is in fact the entire population, the use of RE can be justified by reference to a super-population (all possible observations that could have been) from which the population is 'sampled' (Jones, Owen, et al. 2015).

A difference between the two models which is at least as important as the differences in their assumptions is the set of parameters they estimate. We have already stated that FE models estimate coefficients on higher-level dummy variables (the fixed effects), and cannot estimate coefficients on other higher-level variables (between effects). RE models can yield estimates for coefficients on higher-level variables because the random effects are parameterised as a distribution instead of dummy variables. Moreover, RE automatically provides an estimate of the level 2 variance, allowing an overall measure of the extent to which level 2 units differ in comparison to the level 1 variance. Further, this variance can be used to produce 'shrunken' (or 'Empirical Bayes') versions of the random effects, which take account of the unreliability of those estimates (Bell and Jones 2015). This is often of substantial interest. If we are interested in whether individuals responses are related to their *specific* contexts (neighbourhoods, schools, countries, etc.) a fixed effects model can help answer this question if dummy variables for higher level units are estimated, but this is done unreliably since the unreliability of small higher level units is not accounted for. A RE model can give us a more reliable, appropriately conservative estimate of this, as well as telling us whether that context matters *in general*, based on the size and significance of the estimated variance of the random effects[8]. It can tell us both differences in higher level effects (what in the education literature are termed 'type A' effects - S. W. Raudenbush and Willms (1995) and the effects of variables at the higher level (often 'type B' effects).

The view of FE and RE being defined by their assumptions has led many to characterise the REWB model as a 'hybrid' between FE and RE, or even a 'hybrid FE' model (e.g. see (Schempf et al. 2011). We hope the discussion above will convince readers that this model is a RE model. Indeed, Paul Allison, who (we believe) introduced the terminology of the Hybrid model (Allison 2009) now prefers the terminology of 'within-between RE' (Allison 2014).

Does any of this matter to anyone less pedantic than us? We believe so, or at least it should. First, FE models (and indeed 'hybrid' models) are often presented as a technical solution, following and responding to a Hausman test taken to mean that a RE model cannot be used[9]. As such, researchers rarely consider what specifically is actually being solved, and why the result was wrong in the first place. Indeed, often this bias is described as

[8] This could also be done on the basis of a Wald test of the joint significance of FE dummy variables, although this is a more unwieldy approach, and not possible with non-linear outcomes since conditional likelihood methods do not estimate those dummy variable coefficients.
[9] Many (e.g. Greene, 2012:421) even argue that the model can be used as a form of the Hausman test, which could be itself be used to justify the use of FE, even though the REWB model makes that choice unnecessary.

‘endogeneity’, a term that covers a wide and disparate range of model misspecifications (Bell and Jones, 2015:138). A Hausman test simply investigates whether the between and within effects are different—a possibility that the REWB specification allows for. REWB (a) recognises the possibility of differences between the within and between effects of a predictor, and (b) *explicitly models* those within and between effects simultaneously but separately. King and Roberts (2015) make a similar argument regarding robust standard errors: differences between ordinary and robust standard errors suggest that there is something wrong with the model, such that the next step should be to find out what those misspecifications are. Similarly, we argue that differences between FE and RE models should not be taken to mean the FE model should be used, but rather that something is wrong with the specification of the RE model. The REWB model is a direct, substantive solution to this misspecification in allowing for the possibility of different relations at each level; it models between effects, which may be causing the problem, and are often themselves substantively interesting. When treated as a FE model, this substance is often lost.

Second, using the Hybrid model as if it were a FE model leads researchers to use it without taking full advantage of the benefits that RE models can offer. The RE framework allows a wider range of research questions to be investigated. As stated above, it allows the modelling of time-invariant (or within-group-invariant) variables that are often of substantive interest, and even when they are not of specific interest, can help contextualise the associations of time varying variables. We can easily access reliable, shrunken random effect estimates, which can be of particular interest in some models, for example when the level 2 units are countries and we want to explore the specificities of countries. Where there are more than two levels (for example where individuals observed multiple times are nested within states), all of them can be reflected in a model, and variances at all levels calculated. Furthermore, and as illustrated below, RE models can readily be extended to allow relationships to vary across individuals, or allow variances at any level to vary with variables, that is explicit modelling of heterogeneity (Bullen, Jones, and Duncan 1997). As well as yielding substantively interesting results, such actions can drastically alter the average associations found; it may be, for example, that the 'average effect' found in a standard RE or FE model does not actually apply to any individuals. Under the RE framework, unexplained variances are modelled and parameterised with distributional assumptions, meaning that we can find out more about the world we are studying than if we use a FE framework. Describing REWB, or Hybrid, estimator as falling under a FE framework is therefore misleading about its value and capabilities.[10]

4 Modelling more complexity: random slopes models and three-level models

4.1 Random slopes models

An important limitation of FE models is that, in controlling for differences among clusters, they can hide potentially substantively interesting heterogeneity. The RE/REWB model as previously described can also suffer from this shortcoming, but can also easily avoid it by explicitly modelling such heterogeneity, with the inclusion of random slopes (Western 1998). These allow the coefficients on lower-level covariates to vary across higher-level units. Equation 1 then becomes:

$$y_{it} = \mu + \beta_W(x_{it} - \bar{x}_i) + \beta_B\bar{x}_i + \upsilon_{i0} + \upsilon_{i1}(x_{it} - \bar{x}_i) + \epsilon_{it} \tag{10}$$

Here, υ_{i1} measures the extent to which the average within effect, β_W, varies between individuals. The vectors υ_{i1} and υ_{i0} are assumed to be draws from a bivariate Normal distribution, meaning equation 7 is extended to:

[10] It is also confusing. Walder, Isaacson, and Lu (2015) for example argue, on the basis of a Hausman test, that for their analysis "random-effects models are inappropriate", but they then "estimate the models with a multilevel mixed-effects linear regression" (pp. 456-7).

$$\begin{bmatrix} \upsilon_{i0} \\ \upsilon_{i1} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_{\upsilon 0}^2 & \\ \sigma_{\upsilon 01} & \sigma_{\upsilon 1}^2 \end{bmatrix}\right) \tag{11}$$

Here, the meaning of individual coefficients can vary depending on how variables are scaled and centred. However, the covariance term indicates the extent of 'fanning in' (with negative $\sigma_{\upsilon 01}$) or 'fanning out' (positive $\sigma_{\upsilon 01}$). In many cases, there is substantive heterogeneity in the size of associations among higher-level units. Table 2 shows two such examples of reanalyses where including random coefficients makes a real difference to the results– note that both are analyses of countries, rather than individuals, but the methodological issues are much the same. The first is a reanalysis of an influential study in political science (Milner and Kubota 2005) which claims that democracy leads to globalisation (measured by countries' tariff rates). When including random coefficients in the model, not only does the overall within effect disappear, but a single outlying country, Bangladesh, turns out to be driving the relationship (Bell and Jones 2015), appendix. The second example is the now infamous study in economics by (Reinhart and Rogoff 2010), which claimed that higher levels of public debt cause lower national economic growth. In this case, although the coefficient does not change with the introduction of random slopes, the standard error triples in size, and the within effect is no longer statistically significant when, in addition, time is appropriately controlled.

In both cases, not only is substantively interesting heterogeneity missed in models assuming homogenous associations, but also within effects are anticonservative. Leaving aside the substantive interest that can be gained from seeing how different contexts can lead to different relationships, the message is clear: failing to consider how associations differ across higher-level units can produce misleading results if such differences exist. In particular, using a FE model can lead an analyst to miss problematic assumptions of homogeneity that the model is making. A RE model—including the so-called Hybrid model –

allows for the modelling of important complexities, such as heterogeneity across higher-level units.

We now further demonstrate this important limitation of FE models using a brief simulation study. We simulated data sets with: either 60 groups of 10, or 30 groups of 20; random intercepts distributed Normally, Chi-sq, Normally but with a single large outlier, or with unbalanced groups; with only random intercepts, or both random intercepts and random slopes; and with y either Normal or binary (logit). This produced 32 data-generating processes (DGPs) in total. We then fitted three different models to each simulated dataset: FE, random intercept, and random slope. For the FE models, we calculated both naive and robust SEs.

Table 2: Results from reanalyses of Milner and Kubota 2005, and Reinhart and Rogoff 2010. Standard errors are in parentheses.

Original study/studies	Milner and Kubota 2005	Reinhart and Rogoff 2010, Herndon, Ash, and Pollin 2014
Reanalysis	Bell and Jones 2015 (appendix)	Bell, Johnston, and Jones 2015
Dependent variable	Tariff Rates	Economic Growth (ΔGDP)
Independent variable of interest	Democracy (Polity Score)	National Debt (%GDP)
REWB/FE within estimate (SE)	-0.227 (0.086)**	-0.021 (0.003)***
Random slopes estimate (SE)	-0.143 (0.187) [NS]	-0.021 (0.009)*
Notes	Effect further reduced by the removal of a single outlying country, Bangladesh.	Effect becomes insignificant when time is appropriately controlled.

P values: *** <0.001; ** <0.01; * <0.05; [NS] not significant. For full details of the models used, see the reanalysis papers themselves.

Figure 1 shows the 'optimism'[11] of the standard error on a single covariate, in a variety of scenarios.[12] In the scenarios presented in the top row, the DGP included only random intercepts, not random slopes; the lower row represents DGPs with both random intercepts and random slopes. FE models are in the first two columns (with naïve and robust standard errors), random intercepts models the third column, and random slopes models in the right column.

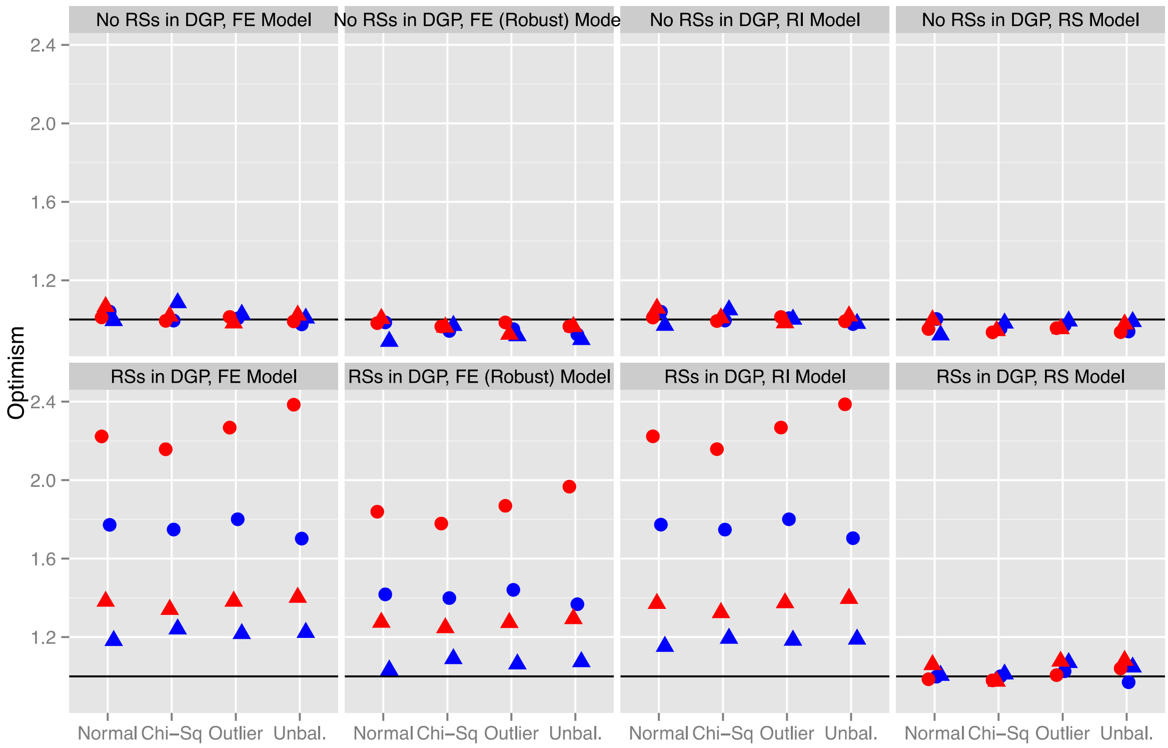
Figure 1 shows that, where random slopes are not included in the analysis model (all but the right-most column), but exist in the data in reality (bottom row), the standard errors are overoptimistic – they are too small relative to the true sampling variability. When there is variation in the slopes across higher level units, there is more uncertainty in the beta estimates, but this is not reflected in the standard error estimates unless those random slopes are explicitly specified. In the top row, in contrast, all four columns look the same: here there is no mismatch between the invariant relationships assumed by the analysis models and present in the data. In the presence of heterogeneity, note that while FE models with naive SEs are the most anticonservative, neither FE models with "robust" standard errors nor RE models with only random intercepts are much better.

These results support the strong critique by Barr et al. (2013) that not to include random slopes is anticonservative. On the other hand, Bates et al. (2015) counter that analytical models should also be parsimonious, and fitting models with many random effects quickly multiplies the number of parameters to be estimated, particularly since random slopes are generally given covariances as well as variances. Sometimes the data available will not be sufficient to estimate such a model. Still, it will make sense in much applied work to test whether a statistically significant coefficient remains so when allowed to vary randomly. We discuss this issue further in the conclusions.

[11] Optimism is the ratio of the true sampling variability to the estimated variability (the latter captured by the standard error—see Shor et al. (2007).

[12] See the appendix for the full R code to replicate these simulations.

Figure 1: Optimism of the Standard Errors in Various Models



Note: Triangles are for logistic models, circles for Normal models; blue means 60 groups of ten and red 30 groups of 20.

4.2 Three (and more) levels, and cross-classifications

Datasets often have structures that span more than two levels; Table 1 provided some examples. A further advantage of the multilevel / random effects framework over fixed effects is its allowing for complex data structures of this kind, which is important when the variable(s) of interest are related to the dependent variable at multiple levels. It is possible to incorporate a within-between distinction into such models (for example, see Raudenbush 2009). Fixed effects models are not problematic when additional higher levels exist (insofar as they can still estimate a within effect), but they are unable to include a third level, because the dummy variables at the second level will automatically use up all degrees of freedom for any levels higher up the hierarchy. Multilevel models allow competing explanations to be considered, specifically at which level in a hierarchy matters the most, with a highly parsimonious specification (by estimating a variance parameter at each level).[13]

For example, cross-national surveys are increasingly being fielded multiple times in the same set of countries, yielding survey data that are both comparative and longitudinal. This presents a three-level hierarchical structure, with observations nested within country-years, which are in turn nested in countries (Fairbrother 2014). However, Schmidt-Catran and Fairbrother (2015) show that in published research, many analyses of such data have failed to include random intercepts at all relevant levels (including 'cross-classified' levels - Goldstein, Browne, and Rasbash 2002), and therefore performed anticonservative tests of key relationships.

4.3 Complex level 1 heterogeneity

A final way in which the random part of the model can be expanded is by allowing the variance at level 1 to vary by one or more covariates. Thus, equation 10 is extended to

$$y_{it} = \mu + \beta_W(x_{it} - \bar{x}_i) + \beta_B\bar{x}_i + \upsilon_{i0} + \upsilon_{i1}(x_{it} - \bar{x}_i) + \epsilon_{it0} + \epsilon_{it1}(x_{it} - \bar{x}_i)\,, \tag{12}$$

where the level 1 variance has two parts, one independent and the other related to $(x_{it} - \bar{x}_i)$. Equation 9 is extended to:

$$\begin{bmatrix}\epsilon_{it0}\\ \epsilon_{it1}\end{bmatrix} \sim N\left(0, \begin{bmatrix}\sigma_{\epsilon 0}^2 & \\ \sigma_{\epsilon 01} & \sigma_{\epsilon 1}^2\end{bmatrix}\right) \tag{13}$$

Often this is important to do, because what is apparent higher level variance between level two units, can in fact be complex variance at level 1. It is only by specifying both, as in

[13] The capability of analysing at multiple scales net of other scales is being exploited in a model- based approach to segregation where the variance at a scale conveys the degree of segregation (Jones, Johnston, et al. 2015).

equation 12, that we can be sure how variance, and varying variance, can be attributed between levels (Vallejo et al. 2015). However, further explication of this is beyond the scope of this paper but we stress its potential importance.

5 Generalising the RE model: binary and count dependent variables

So far, this paper has considered only models with continuous dependent variables, using an identity link function. But do the claims of this paper apply to Generalised Linear models? These include other dependent variables and link functions (Neuhaus and McCulloch 2006), such as logit models (for binary/proportion dependent variables) and Poisson models (for count dependent variables). Although this question has not been considered to a great extent in the social and political sciences, the biostatistics literature does provide some answers (for an accessible discussion of this, see Allison 2014). Here we briefly outline some of the issues, although as this is not the focus of the paper, so we do not address all.

Unlike models using the identity link function, results using the REWB model with other link functions do not produce results that are identical to FE (or the equivalent conditional likelihood model). In other words, the inclusion of the group mean in the model does not reliably partition any higher level processes from the within effect, meaning both within and between estimates can be biased. This is the case when the relationship between the between component of X ($\bar{x}_i$) and the higher level residual (υ_i) is non-linear. How big a problem is this? Brumback et al. (2010:1651) found that, in running simulations, "it was difficult to find an example in which the problem is severe" (see also Goetgeluk and Vansteelandt 2008). In a later paper, however, Brumback, Zheng, and Dailey (2013) did identify one such example, but only with properties unlikely to be found in real life data (Allison 2014) - $\bar{x}_i$ and υ_i very highly correlated, and few observations per unit.

Whether the REWB model should be used, or a conditional likelihood (FE) model should be used instead, depends on three factors: (1) the link function, (2) the nature of the research question, and (3) the researcher's willingness to accept low levels of bias. Regarding (1), many link functions, including negative binomial models, ordered logit models, and probit models, do not have a conditional logit associated with them. If such models are to be used, the REWB model may be the best method available to produce within effects that are (relatively) unbiased by omitted higher-level variables. Regarding (2), conditional likelihood methods have all the disadvantages of FE mentioned above: they are unable to provide cluster-level effects, random slopes cannot be fitted, and so on, meaning there is a risk of producing misleading and anti-conservative results. These will often be important to the research question at hand, to provide a realistic level of complexity to the modelling of the scenarios at hand. Finally, relating to point (3), the level of bias is easily ascertained by comparing the estimate of the REWB model to that of the conditional likelihood model (where available). If the results are deemed similar enough, the researcher can be relatively sure that the results produced by the REWB model are reasonable.

6 Assumptions of random effects models: how much do they matter?

A key assumption of RE models is that the random effects representing the higher level entities are drawn from a Normal distribution. However, "the Normality of [the random coefficients] is clearly an assumption driven more by mathematical convenience than by empirical reality" (Beck and Katz 2007:90). Indeed, it is often an unrealistic assumption, and it is important to know the extent to which different estimates are biased when that assumption is broken.

The evidence from prior simulations studies is somewhat mixed, and depends on what specifically in the RE model is of interest. For continuous-Y linear models, and on the positive side, Beck and Katz (2007) find that both average parameter estimates and random effects are well estimated, both when the random effects are assumed to be Normally distributed but are in fact distributed by a chi-squared distribution, or there are a number of outliers in the dataset.[14] However, Beck and Katz's simulations only considered the performance of the model in relation to OLS single level regression, and only considered root mean squared error (RMSE), which combines bias and efficiency; they therefore did not consider the absolute magnitude of bias. Others concur that beta estimates are generally unbiased by non-Normal random effects, as are estimates of the RE variances (Maas and Hox 2004; McCulloch and Neuhaus 2011a). Random effects are only biased to a significant degree in extreme scenarios (McCulloch and Neuhaus 2011b), and even then (for example for random effects with a Chi-sq(1) distribution), the ranked order of estimated random effects remains highly correlated (Corr > 0.8) to the rankings of the true random effects (Arpino and Varriale 2010), meaning substantive interpretation is likely to be affected only minimally. This is the case whether or not the DGP includes random slopes. In other words, a badly specified random distribution may result in some biases, but these are usually small enough not to worry the applied researcher. If there is a worry about bias, it may be wise to check the findings are robust to other specifications, and potentially use models that allow for non-Normal random effects, such as Non-Parametric Maximum Likelihood techniques (Aitkin 1999; Fotouhi 2003).

With non-linear models, the evidence is somewhat less positive. Where the Normality assumption of the higher-level variance is violated, there can be significant biases, particularly on level 2 estimators when the level 2 variance is large (as is often the case with panel data, but not in cross-sectional data (Heagerty and Kurland 2001). For a review of these simulation studies, see (Grilli and Rampichini 2015).

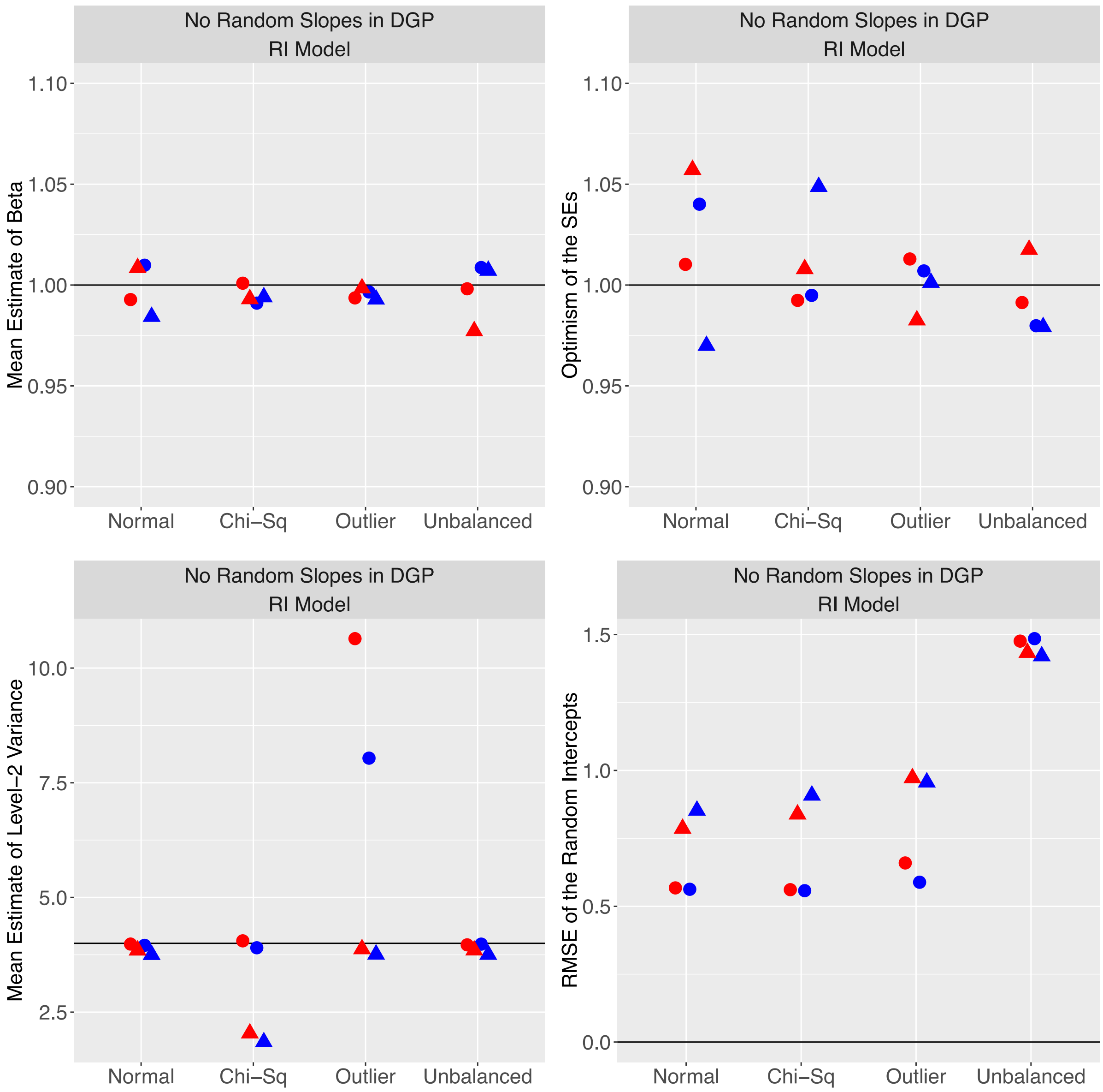
[14] In the latter case, outlying random effects can easily be identified and 'dummied out', allowing the distribution of the rest of the random effects to be estimated.

Our simulations, for the most part, back up these findings and this is illustrated in Figure 2, which presents the consequences for various parameters if the random intercepts are distributed Chi-sq(2), or have a single substantial outlier, and if the groups are unbalanced. First, beta estimates are unbiased (upper-left panel), as are their standard errors (upper-right), regardless of the true distribution of the random effects and the type of model.

Non-Normality does however have consequences for the estimate of the level-2 variance (lower-left panel). When the true distribution is skewed (in a Chi-squared(2) distribution), for logistic models there is notable downward bias in the estimate of the level two variance, and a slight increase in the error associated with the random effects themselves (lower-right). We find no evidence of any similar bias with continuous-Y models, however. In contrast, when the non-Normality of the random effects is due to an outlying level-2 entity, there is an impact on the estimated variance for continuous-Y models, and the estimated random intercepts for both logistic and Normal models. However, as noted above, the latter does not need to be problematic, because outliers can be easily identified and 'dummied out', effectively removing that specific random effect from the estimated distribution. Note that the high RMSE associated with unbalanced datasets (lower-right) is related to the smaller sample size in some level 2 groups, rather than being evidence of any bias.

In sum, even substantial violations of the Normality assumption of the higher-level random effects do not have much impact on beta estimates in the fixed part of the model, nor the standard errors. Such violations can however affect the random effects estimates, particularly in the case of non-continuous-Y models.

Figure 2: Biases and RMSE under Various (Mis-)Specifications



Note: Triangles are for logistic models, circles for Normal models; blue means 60 groups of ten—red 30 groups of 20. Clockwise from the upper-left, the parameters are beta (bias), optimism of the standard errors (bias), random intercepts (RMSE), and level-2 variance (bias).

7 Conclusion: what should researchers do?

We hope that this article has presented a picture of the relative merits of fixed effects and random effects models, including REWB models. We have considered what each of these models are, what they do, what they assume, and how much those assumptions matter in different real life scenarios. There are of course other statistical issues with these models that we have not addressed (for example, the issue of lags: see Vaisey and Miles (2014) for a good discussion of this). However such issues can be incorporated into either FE or RE modelling (Spencer 2003; Steele 2007), so it makes sense to keep them separate from this discussion.

There are a number of practical points that researchers should take away from this paper. First and perhaps most obviously is that in many if not most scenarios, the REWB model is a better more encompassing modelling option than either FE or conventional RE, which do not distinguish between within and between effects. Even when using non-identity link functions, or when the Normality assumption of the random effects is violated, the biases that can arise in such models will often be a price worth paying for the added flexibility that the REWB model provides. This is especially the case since FE is unable to provide estimates at all of the parameters that are most biased by violations of Normality (specifically random effects and variance estimates).

Second, the question of whether to include random slopes is important and requires careful consideration. On the one hand, in a world of limited computing power and limited data, it is not feasible to allow the effects of all variables to vary between units. On the other hand, we have shown that results can change in substantive ways when slopes are allowed to vary randomly. We would argue that, at the least, where there is a single substantive variable of interest, it would make sense to check that the conclusions hold when the effect of that variable is allowed to vary. One option in this regard is to use robust standard errors, not as a correction per se, but as a diagnostic procedure – a 'canary down the mine' – following King and Roberts (2015). Any difference between classical and robust standard errors suggests there is some kind of misspecification in the model, and that misspecification might well include the failure to model random slopes.

Third, and in contravention to much of the applied literature, we argue that researchers should not use a Hausman test to decide between fixed and random effects models. Rather, they can use this test, or models equivalent to it, to verify the equivalence of the within and between effects. A lack of equality should be in itself of interest and worthy of further investigation through the within-between model, rather than a sign that anything needs to be 'corrected' per se.

8 Glossary

This is not a complete list; rather it focuses on terms that are often confused, or have multiple meanings. We refer readers to another glossary (Diez Roux 2002) for further definitions of terms.

Between effects – the association between the higher level average level of an independent variable and the dependent variable. The actual value of that variable (at level 1) is not controlled, and for this reason is often more of interest with panel data (e.g. occasions nested in individuals) than in cross sectional data (e.g. individuals nested in neighbourhoods) (see section 2.1). If a causal interpretation is desired, the estimate could be subject to omitted variable bias as a result of omitted variables at level 2 (see section 2.3).

Conditional Likelihood estimation – a collection of models that control out cluster level effects. The fixed effects model, as described in this paper, is an example of a conditional likelihood model, specifically for models with a continuous dependent variable that use the identity link function. Other conditional likelihood methods exist for other types of model / dependent variable, including logit models (for binary dependent variables) and Poisson models (for count data). They do not exist for other link functions, including negative binomial models, probit models, and ordered logit models.

Contextual effects – the association between the higher level average level of an independent variable and the dependent variable, controlling for the actual level of that variable (at level 1). It is therefore more of interest with cross sectional data (e.g. individuals nested in neighbourhoods) than with panel data (e.g. occasions nested in individuals) (see section 2.1). If a causal interpretation is desired, the estimate could be subject to omitted variable bias as a result of omitted variables at level 2 (see section 2.3).

Endogeneity (1) – In regression we assume that the independent variables cause the dependent variable, and thus that there is no internal cause of the X variable. If in fact the direction of causality is the reverse, then that X variable can be considered endogenous.

Endogeneity (2) – when independent variables are correlated with the error term. By this definition, there can be many different causes: endogeneity as described above, omitted variable bias, selection bias, and so on. It is therefore not a particularly useful term by this definition.

Fixed Effects (1) – the method – a model that includes dummy variables for higher level entities, or performs some transformation to condition out those entities. Thus, no assumptions are made about the higher-level entities, because they are no longer present in the model. Equally, nothing can be said about those higher level entities, or any variable that is measured at that level.

Fixed Effects (2) – the coefficients associated with the dummy variables themselves.

Fixed part (of a model) – a part of a multilevel model where the terms being estimated are fixed to a single value across the sample, including the overall intercept and slope coefficients.

Hausman test – a test to compare the estimates of two different models, most commonly a fixed effects and a random effects model. A difference between the two should provide evidence that the within and between effects of one or more variables are not the same, and therefore that there are level 2 omitted variables biasing the estimated effects of level 1 variables. It should not be taken to mean that the fixed effects model should be used, however, since this issue can be solved by including the group mean of the variable in the model (either with the Mundlak model or the within-between random effects (REWB) model.

Hybrid Model – a rather misleading term for either the within-between random effects (REWB) model (or the Mundlak model). It is misleading because it implies that the models are part fixed effects and part random effects, when in fact they are unequivocally random effects models.

Level 1 – the lower level. In panel data, this represents the occasion. With cross sectional data it represents the individual

Level 2 – the higher level. In panel data, this represents the individual (or country). With cross sectional data it might represent the neighbourhood, country etc.

Level 3 – for models with a more complex structure, there may be more than two levels. For example, in a panel that is repeated across countries, you might have occasions (level 1) nested in individuals (level 2) nested in countries (level 3).

Mundlak Model – a random effects model that includes the raw variables and the group means of the variables. The coefficients associated with the raw variables are estimates of within effects, and will be equivalent to estimates from a fixed effects model – i.e. the problem of level 2 omitted variables is solved for these coefficients. The coefficients associated with the group means will be contextual effects

Omitted Variables – variables that are not included in your analysis. They must therefore be considered as potential reasons for estimated model coefficients. In these cases, it would be wrong to assume that a particular coefficient represents a causal effect.

Omitted Variable Bias – the result of omitting important variables from an analysis, when those variables are correlated with both another independent variable and Y. However this is only really bias if the association estimated is interpreted as if that omitted variable didn't exist. For example in education research, when interpreting the association between 'free school meal' eligibility and school performance, we interpret it correctly as an indicator of a proxy for an omitted variable, income. If we interpreted it as the effect of the school meals themselves, we would be ignoring the omitted variable (income) and our interpretation would be problematic.

Poor man's conditional likelihood – a rather unfair, in our opinion, name for the Within-between random effects model, used by Neuhaus and McCulloch (2006) when applied to the generalised linear mixed model.

Random effects (1) – the method – a model that assumes higher entities are a sample that comes from a Normal distribution, the variance of which is estimated. If this assumption is broken, there is limited bias on estimated coefficients, at least for fixed effects estimates and variance estimates. However, violation of this assumption has more of an effect in non-Normal models.

Random effects (2) – the higher level entities that make up the sample of the distribution. These can be either the raw random effects, or those that have been shrunk to account for the uncertainty in each estimate. The estimated variance of these random effects is based on the raw random effects.

Random slopes model – an extension of a random effects model, where the association of a level 1 variable is allowed to vary across level 2 units. Also called a random coefficient model,

Random part (of a model) – the part of a multilevel model that is allowed to vary – that is the variance parameters and the random effects.

Within-between random effects model (REWB) – a random effects model de-meaned variables, and the means of those variables. The coefficients associated with the de-meaned variables are estimates of within effects, and will be equivalent to estimates from a fixed effects model – i.e. the problem of level 2 omitted variables is solved for these coefficients. The coefficients associated with the variable means are between effects.

Within effects – the association between the level of an independent variable *relative to its average value for that higher level unit*, and the dependent variable. When interpreted causally, it is the effect of an unusually high (or low) level of an independent variable, compared to the average for that higher level unit. It could be subject to omitted variable bias as a result of omitted variables at level 1, but not at level 2.

9 References

Aitkin, Murray. 1999. "A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models." *Biometrics* 55(1): 117–28.

Allison, Paul D. 1994. "Using Panel Data to Estimate the Effects of Events." *Sociological Methods & Research* 23(2): 174–99.

———. 2009. *Fixed Effects Regression Models*. London: Sage.

———. 2014. "Problems with the Hybrid Method." *Statistical Horizons*. http://www.statisticalhorizons.com/problems-with-the-hybrid-method (July 16, 2015).

Arpino, Bruno, and Roberta Varriale. 2010. "Assessing the Quality of Institutions' Rankings Obtained through Multilevel Linear Regression Models." *Journal of Applied Economic Sciences* 5(1): 7–22.

Beck, Nathaniel, and Jonathan N Katz. 1995. "What to Do (and Not to Do) with Time-Series Cross-Section Data." *American Political Science Review* 89(3): 634–47.

———. 2007. "Random Coefficient Models for Time-Series-Cross-Section Data: Monte Carlo Experiments." *Political Analysis* 15(2): 182–95.

Bell, Andrew, Ron Johnston, and Kelvyn Jones. 2015. "Stylised Fact or Situated Messiness? The Diverse Effects of Increasing Debt on National Economic Growth." *Journal of Economic Geography* 15(2): 449–72.

Bell, Andrew, and Kelvyn Jones. 2015. "Explaining Fixed Effects: Random Effects Modelling of Time-Series Cross-Sectional and Panel Data." *Political Science Research and Methods* 3(1): 133–53.

Brumback, Babette A et al. 2010. "Adjusting for Confounding by Cluster Using Generalized Linear Mixed Models." *Statistics & Probability Letters* 80(21-22): 1650–54.

Brumback, Babette A, Hao W Zheng, and Amy B Dailey. 2013. "Adjusting for Confounding by Neighborhood Using Generalized Linear Mixed Models and Complex Survey Data." *Statistics in medicine* 32(8): 1313–24.

Bullen, Nina, Kelvyn Jones, and Craig Duncan. 1997. "Modelling Complexity: Analysing between-Individual and between-Place Variation - A Multilevel Tutorial." *Environment and Planning A* 29(4): 585–609.

Chatelain, Jean-Bernard, and Kirsten Ralf. 2010. "Inference on Time-Invariant Variables Using Panel Data: A Pre-Test Estimator with an Application to the Returns to Schooling." *PSE Working Paper*. http://hal-paris1.archives-ouvertes.fr/docs/00/49/20/39/PDF/Chatelain_Ralf_Time_Invariant_Panel.pdf (March 28, 2016).

Clark, Tom S, and Drew A Linzer. 2015. "Should I Use Fixed or Random Effects?" *Political Science Research and Methods* 3(2): 399–408.

Deeming, Christopher, and Kelvyn Jones. 2015. "Investigating the Macro Determinants of Self-Rated Health and Well-Being Using the European Social Survey: Methodological Innovations across Countries and Time." *International Journal of Sociology* 45(4): 256–85.

Delgado-Rodríguez, Miguel, and Javier Llorca. 2004. "Bias." *Journal of Epidemiology and Community Health* 58(8): 635–41.

Diez Roux, A V. 2002. "A Glossary for Multilevel Analysis." *Journal of Epidemiology & Community Health* 56(8): 588–94.

Fairbrother, Malcolm. 2014. "Two Multilevel Modeling Techniques for Analyzing Comparative Longitudinal Survey Datasets." *Political Science Research and Methods* 2(1): 119–40.

Fotouhi, Ali Reza. 2003. "Comparisons of Estimation Procedures for Nonlinear Multilevel Models." *Journal of Statistical Software* 8(9): 1–39.

Gelman, Andrew. 2005. "Why I Don't Use the Term 'Fixed and Random Effects.'" *Statistical Modeling, Causal Inference, and Social Science*. http://andrewgelman.com/2005/01/25/why_i_dont_use/ (November 19, 2015).

Goetgeluk, Sylvie, and Stijn Vansteelandt. 2008. "Conditional Generalized Estimating Equations for the Analysis of Clustered and Longitudinal Data." *Biometrics* 64(3): 772–80.

Goldstein, Harvey, William J Browne, and Jon Rasbash. 2002. "Multilevel Modelling of Medical Data." *Statistics in medicine* 21(21): 3291–3315.

Grilli, Leonardo, and Carla Rampichini. 2015. "Specification of Random Effects in Multilevel Models: A Review." *Quality & Quantity* 49(3): 967–76.

Halaby, Charles N. 2004. "Panel Models in Sociological Research: Theory into Practice." *Annual Review of Sociology* 30(1): 507–44.

Hanchane, Saïd, and Tarek Mostafa. 2012. "Solving Endogeneity Problems in Multilevel Estimation: An Example Using Education Production Functions." *Journal of Applied Statistics* 39(5): 1101–14.

Hausman, Jerry A. 1978. "Specification Tests in Econometrics." *Econometrica* 46(6): 1251–71.

Heagerty, P. J., and B. F. Kurland. 2001. "Misspecified Maximum Likelihood Estimates and Generalised Linear Mixed Models." *Biometrika* 88(4): 973–85.

Herndon, Thomas, Michael Ash, and Robert Pollin. 2014. "Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff." *Cambridge Journal of Economics* 38(2): 257–79.

Howard, A. L. 2015. "Leveraging Time-Varying Covariates to Test Within- and Between-Person Effects and Interactions in the Multilevel Linear Model." *Emerging Adulthood* 3(6): 400–412.

Jones, Kelvyn, Ron Johnston, et al. 2015. "Ethnic Residential Segregation: A Multilevel, Multigroup, Multiscale Approach Exemplified by London in 2011." *Demography* 52(6): 1995–2019.

Jones, Kelvyn, Dewi Owen, et al. 2015. "Modelling the Occupational Assimilation of Immigrants by Ancestry, Age Group and Generational Differences in Australia: A Random Effects Approach to a Large Table of Counts." *Quality & Quantity* 49(6): 2595–

2615.

King, Gary, and Margaret Roberts. 2015. "How Robust Standard Errors Expose Methodological Problems They Do Not Fix." *Political Analysis* 23(2): 159–79.

Kloosterman, R., N. Notten, J. Tolsma, and G. Kraaykamp. 2010. "The Effects of Parental Reading Socialization and Early School Involvement on Children's Academic Performance: A Panel Study of Primary School Pupils in the Netherlands." *European Sociological Review* 27(3): 291–306.

Lauen, Douglas Lee, and S. Michael Gaddis. 2013. "Exposure to Classroom Poverty and Test Score Achievement: Contextual Effects or Selection?" *American Journal of Sociology* 118(4): 943–79.

Maas, Cora J M, and Joop J Hox. 2004. "Robustness Issues in Multilevel Regression Analysis." *Statistica Neerlandica* 58(2): 127–37.

Maimon, David, and Danielle C Kuhl. 2008. "Social Control and Youth Suicidality: Situating Durkheim's Ideas in a Multilevel Framework." *American Sociological Review* 73(6): 921–43.

McCulloch, Charles E, and John M Neuhaus. 2011a. "Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter." *Statistical Science* 26(3): 388–402.

———. 2011b. "Prediction of Random Effects in Linear and Generalized Linear Models under Model Misspecification." *Biometrics* 67(1): 270–79.

Milner, Helen V, and Keito Kubota. 2005. "Why the Move to Free Trade? Democracy and Trade Policy in the Developing Countries." *International Organization* 59(1): 107–43.

Mundlak, Yair. 1978. "Pooling of Time-Series and Cross-Section Data." *Econometrica* 46(1): 69–85.

Neuhaus, John M, and Charles E McCulloch. 2006. "Separating between- and within-Cluster Covariate Effects by Using Conditional and Partitioning Methods." *Journal of the Royal Statistical Society Series B-Statistical Methodology* 68: 859–72.

Rasbash, Jon. 2008. "Module 4: Multilevel Structures and Classifications." *LEMMA VLE*. http://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/4-concepts-sample.pdf (November 19, 2015).

Raudenbush, Stephen W. 2009. "Adaptive Centering with Random Effects: An Alternative to the Fixed Effects Model for Studying Time-Varying Treatments in School Settings." *Education, Finance and Policy* 4(4): 468–91.

Raudenbush, Stephen W, and Anthony Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*. 2nd ed. London: Sage.

Raudenbush, Stephen W, and J. Willms. 1995. "The Estimation of School Effects." *Journal of Educational and Behavioral Statistics* 20(4): 307–35.

Reinhart, Carmen M, and Kenneth S Rogoff. 2010. "Growth in a Time of Debt." *American Economic Review* 100(2): 573–78.

Ruiter, Stijn, and Frank van Tubergen. 2009. "Religious Attendance in Cross-National Perspective: A Multilevel Analysis of 60 Countries." *American Journal of Sociology* 115(3): 863–95.

Sampson, R. J., Stephen W Raudenbush, and Felton Earls. 1997. "Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy." *Science* 277(5328): 918–24.

Schempf, Ashley H, Jay S Kaufman, Lynne C Messer, and Pauline Mendola. 2011. "The

Neighborhood Contribution to Black-White Perinatal Disparities: An Example from Two North Carolina Counties, 1999-2001." *American journal of epidemiology* 174(6): 744–52.

Schmidt-Catran, Alexander W., and Malcolm Fairbrother. 2015. "The Random Effects in Multilevel Models: Getting Them Wrong and Getting Them Right." *European Sociological Review* 32(1): 23–38.

Schmidt-Catran, Alexander W., and D. C. Spies. 2016. "Immigration and Welfare Support in Germany." *American Sociological Review* online.

Shor, Boris, Joseph Bafumi, Luke Keele, and David Park. 2007. "A Bayesian Multilevel Modeling Approach to Time-Series Cross-Sectional Data." *Political Analysis* 15(2): 165–81.

Snijders, Tom A B, and Roel J Bosker. 2012. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling*. 2nd ed. London: Sage.

Spencer, Neil H. 2003. "Consistent Parameter Estimation for Lagged Multilevel Models." *Journal of Data Science* 1(2): 123–47.

Spiegelhalter, David J. 2004. "Incorporating Bayesian Ideas into Health-Care Evaluation." *Statistical Science* 19(1): 156–74.

Steele, Fiona. 2007. "Multilevel Models for Longitudinal Data." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 171(1): 5–19.

Steele, Fiona, Anna Vignoles, and Andrew Jenkins. 2007. "The Effect of School Resources on Pupil Attainment: A Multilevel Simultaneous Equation Modelling Approach." *Journal of the Royal Statistical Society Series a-Statistics in Society* 170: 801–24.

Subramanian, S V, Kelvyn Jones, Afamia Kaddour, and Nancy Krieger. 2009. "Revisiting Robinson: The Perils of Individualistic and Ecologic Fallacy." *International Journal of Epidemiology* 38(2): 342–60.

Vaisey, Stephen, and Andrew Miles. 2014. "What You Can - and Can't - Do with Three-Wave Panel Data." *Sociological Methods & Research* online: 1–24.

Vallejo, Guillermo, Paula Fernández, Marcelino Cuesta, and Pablo E. Livacic-Rojas. 2015. "Effects of Modeling the Heterogeneity on Inferences Drawn from Multilevel Designs." *Multivariate Behavioral Research* 50(1): 75–90.

Walder, A. G., A. Isaacson, and Q. Lu. 2015. "After State Socialism: The Political Origins of Transitional Recessions." *American Sociological Review* 80(2): 444–68.

Western, Bruce. 1998. "Causal Heterogeneity in Comparative Research: A Bayesian Hierarchical Modelling Approach." *American Journal of Political Science*: 1233–59.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge MA: MIT Press.