Explaining Fixed Effects: Random Effects modelling of Time-Series Cross-Sectional and Panel Data

Andrew Bell and Kelvyn Jones

School of Geographical Sciences

Centre for Multilevel Modelling

University of Bristol

Last updated: 11th Sept 2013

Draft – please do not cite without permission

Contact: andrew.bell@bristol.ac.uk

Contents

Abstract

This article challenges Fixed Effects (FE) modelling as the 'default' for time-series-cross-sectional and panel data. Understanding differences between within- and between-effects is crucial when choosing modelling strategies. The downside of Random Effects (RE) modelling – correlated lower-level covariates and higher-level residuals – is omitted-variable bias, solvable with Mundlak's (1978a) formulation. Consequently, RE can provide everything FE promises and more, and this is confirmed by Monte-Carlo simulations, which additionally show problems with another alternative, Plümper and Troeger's Fixed Effects Vector Decomposition method, when data are unbalanced. As well as being able to model time-invariant variables, RE is readily extendable, with random coefficients, cross-level interactions, and complex variance functions. An empirical example shows that disregarding these extensions can produce misleading results. We argue not simply for technical solutions to endogeneity, but for the substantive importance of context and heterogeneity, modelled using RE. The implications extend beyond political science, to all multilevel datasets.

Acknowledgements

Thanks to Helen Milner and Keito Kubota for making their data available to us, and to Fiona Steele, Paul Clarke, Malcolm Fairbrother, Alastair Leyland, Mark Bell, Ron Johnston, George Leckie, Dewi Owen, Nathaniel Beck, Chris Adolph, and Thomas Plümper for their help and advice. Also, thanks to the two anonymous reviewers for their suggestions. None of these are responsible for what we have written.

Keywords

Random Effects models, Fixed Effects models, Random coefficient models, Mundlak formulation, Fixed effects vector decomposition, Hausman test, Endogeneity, Panel Data, Time-Series Cross-Sectional Data.

1 Introduction

Two solutions to the problem of hierarchical data, with variables and processes at both a higher and lower-level, vie for prominence in the social sciences. Fixed effects (FE) modelling is used more frequently in economics and political science reflecting its status as the "gold standard" default (Schurer and Yong, 2012 p1). However Random effects (RE) models, also called multilevel models, hierarchical linear models, and mixed models, have gained increasing prominence in political science (Beck and Katz, 2007), and are used regularly in education (O'Connell and McCoach, 2008), epidemiology (Duncan et al., 1998), geography (Jones, 1991) and biomedical sciences (Verbeke and Molenberghs, 2000, 2005). Both methods are applicable to research questions with complex structure, including both place-based hierarchies [such as individuals nested within neighbourhoods, for example Jones et al. (1992)], and temporal hierarchies [such as panel data and time-series cross-sectional (TSCS) data[1], where measurement occasions are nested within entities such as individuals or countries (see Beck, 2007)]. Whilst this article is particularly concerned with the latter, its arguments apply equally to all forms of hierarchical data[2].

One problem with the disciplinary divides outlined above is that much of the debate between the two methods has remained separated by subject boundaries, with the two sides of the debate seeming to often talk past each other. This is a problem, because we believe that both sides are making important points which are currently not taken seriously

[1] The difference between TSCS and Panel data lies partly in its sample structure: TSCS data has comparatively few higher level entities (usually groups of individuals such as countries, rather than individuals) and comparatively many measurement occasions (Beck and Katz, 1995). In addition, TSCS data, used mainly in political science, often contains more slowly changing, historically determined variables (such as GDP per capita) and researchers using it are often more interested in specific effects in specific higher-level entities. This makes the issues we discuss here particularly important to researchers using TSCS data.

[2] Indeed, this includes non-hierarchical data with cross-classified or multiple membership structures (see Snijders and Bosker 2012 p205).

by their counterparts. This article draws on a wide, multidisciplinary literature and as such we hope that it will go some way towards informing each side of the relative merits of both sides of the argument.

Having said this, we take the strong and rather heterodox view that there are few, if any, occasions in which FE modelling is preferable to RE modelling. If the assumptions made by RE models are correct, RE would be the preferred choice because of its greater flexibility and generalisability, and its ability to model context, including variables that are only measured at the higher level. We show in this article that the assumptions made by RE models, including the exogeneity of covariates and the Normality of residuals, are at least as reasonable as those made by FE models when the model is correctly specified. Unfortunately, this correct formulation is used all too rarely (Fairbrother, 2011) despite being fairly well known [it is discussed in numerous econometrics textbooks (Greene, 2012, Wooldridge, 2002), if rather too briefly]. Furthermore, we argue that, in controlling out context, FE models effectively cut out much of what is going on, goings-on which are usually of interest to the researcher, the reader, and the policy maker. Models which control out, rather than explicitly model, context and heterogeneity offer overly simplistic and impoverished results which can lead to misleading interpretations.

This article's title has two meanings. First, we hope to explain the technique of fixed effects estimation to those who use it too readily as a default option without fully understanding what they are estimating and what they are losing by doing so. And second, we show that whilst the fixed dummy coefficients in the FE model are measured unreliably, RE models are able to explain and thus reveal specific differences between higher-level entities.

The distinctive features of this paper are as follows[3]:

- It is a central attack on the dominant method in much of the quantitative social sciences, and as such makes a much more forceful argument against FE modelling than has been made before. We see the FE model as a special and rather restricted case of the appropriately formulated RE model.
- It argues for an alternative approach to endogeneity: a concern for its causes, which in this case is separate and potentially different 'within' and 'between' effects, that need to be studied, thought about and modelled explicitly, rather than for it simply to be eliminated with little regard for what is being lost in the process[4].
- It emphasises the importance of explicitly modelling heterogeneity, and not just mean effects. It argues that implementing this in a RE framework is often essential and that failing to do so can lead to incorrect inference. It thus extends the basic method suggested here and by others (Bafumi and Gelman, 2006, Bartels, 2008) through random coefficients and cross-level interactions.

The article proceeds as follows. We first outline a basic RE model, and show its comparative advantages over other oft-used techniques. We then outline what is undoubtedly a critical problem with many RE models: correlation between lower-level predictors and higher-level residuals, and show why this 'endogeneity' occurs so regularly, alongside consideration of the much misunderstood Hausman specification test (Hausman, 1978). We then outline the FE solution which circumvents this problem by controlling out all differences between

[3] The paper does not address issues of dynamics; however there is no reason why existing methods for accounting for dynamic effects could not be incorporated into models like those suggested here. For example, Zorn (2001) suggest separating within and between effects in a discrete-time duration model for dyadic data.

[4] In this we are following Hanchane and Mostafa (2011) who show that different levels of endogeneity in education production functions are produced by the context of a country's educational system and are of substantive interest.

higher level units. However, the FE model is very limited in being unable to estimate the effects of higher-level variables; this is discussed in the following section, alongside a FE-based solution proposed by Plümper and Troeger (2007). We then show that, in fact, this solution has many characteristics of a RE model, and that the latter, used with a formulation similar to that originally proposed by Mundlak (1978a) which partitions the effect of lower-level covariates into two parts, is a more parsimonious and flexible method for achieving the same thing. This solution treats endogeneity as a substantive phenomenon, which occurs when a given lower-level variable with different within and between processes is assumed to have a single homogenous effect. The efficacy of this model, at least in comparison to the suggested alternatives, is shown by Monte-Carlo simulations. This is followed by consideration of an extension of this model that lets additional coefficients vary randomly, allowing for cross-level interactions and the estimation of variance functions, and finally an example which shows that failing to implement these extensions can lead to very misleading results. It is important to say once again that our recommendations are not entirely one-sided: the formulation that we propose is currently not used enough[5] and in many disciplines endogeneity is often ignored. However the point remains: a well-specified RE model can be used to achieve everything that FE models achieve, and much more besides.

[5] Endogeneity is notable in its absence from multilevel modelling quality checklists (such as Ferron et al., 2008). Indeed, the following Google scholar 'hits' of combinations of terms (24th April 2012) tells their own story:

Terms	With "Hausman"	With "Mundlak"
"Fixed effects"	25,000	1960
"Random effects"	18,900	1610
"Multilevel"	2,400	170

The multilevel modelling literature has not significantly engaged with the Mundlak formulation or the issue of endogeneity.

2 The problem of hierarchies in data, and the Random Effects solution

Many research problems in the social sciences have a hierarchical structure; indeed "once you know hierarchies exist, you see then everywhere" (Kreft and De Leeuw, 1998 p1). Such hierarchies are produced because the population is hierarchically structured – voters at level 1 are nested in constituencies at level 2 – and/or a hierarchical structure is imposed during data collection so that, for example in a longitudinal panel, there are repeated measures at level 1 nested in individuals at level 2. In the discussion that follows and to make things concrete we use 'higher-level entities' to refer to level 2, and occasions to refer to level 1. Consequently, time-varying observations are measured at level 1 and time-invariant observations at level 2; the latter are unchanging attributes. Thus, in a panel study, higher-level entities are individuals, and time-invariant variables may include characteristics such as gender. In a TSCS analysis, the higher-level entities may be countries, and time-invariant variables could be whether they are located in the global south.[6]

The technical problems of the analysis of hierarchies in data are well known. Put briefly, standard 'pooled' linear regression models assume that residuals are independently and identically distributed (IID). That is, once all covariates are considered, there are no further correlations (i.e. dependence) between measures. Substantively, this means that the model assumes that any two higher-level entities are identical and thus they can be completely 'pooled' into a single population. With hierarchical data, particularly with temporal hierarchies which are often characterised by marked dependence over time, this is patently an unreasonable assumption. Responses for measurement occasions within a given higher-level entity are often related to each other. As a result, the effective sample size of such

[6] Duncan, et al. (1998) develop this perspective whereby a range of research questions and different research designs are seen as having hierarchical or more complex structure.

datasets is much smaller than a simple regression would assume: closer to the number of higher-level entities (individuals, or countries) than the number of lower-level units (measurement occasions). As such standard errors will be incorrect[7] if this dependence is not taken into account (Moulton, 1986).

The RE solution to this dependency is to partition the unexplained residual variance into two: higher-level variance between higher level entities and lower-level variance within these entities, between occasions. This is achieved by having a residual term at each level, the higher level residual being the so-called random effect. As such a simple standard RE model would be:

$$y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + e_{ij}$$

where

$$\beta_{0j} = \beta_0 + \beta_2 z_j + u_j.$$

These are the 'micro' and 'macro' parts of the model respectively and they are estimated together in a combined model which is formed by substituting the latter into the former:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 z_j + (u_j + e_{ij}) \quad (1)$$

where y_{ij} is the dependent variable. In the 'fixed part' of the model β_0 is the intercept term, x_{1ij} is a (series of) covariate(s) which are measured at the lower, occasion level with coefficient β_1, and z_j is a (series of) covariate(s) measured at the higher level with coefficient β_2. The 'random part' of the model (in brackets) consists of u_j, the higher-level residual for higher-level entity j, allowing for differential intercepts for higher-level entities,

[7] Standard errors will usually be underestimated in pooled OLS which ignores the hierarchical structure, but can also be biased up (see Arceneaux and Nickerson, 2009 p185).

and e_{ij}, the occasion-level residual for occasion i of higher-level entity j. The u_j term is in effect a measure of 'similarity' that allows for dependence as it applies to all the repeated measures of a higher-level entity. The variation that occurs at the higher level (including u_j and any time-invariant variables) is considered in terms of the (smaller) higher-level entity sample size, meaning that the standard errors are correct. By assuming that u_j and e_{ij} are Normally distributed, an overall measure of their respective variances can be estimated:

$$u_j \sim N(0, \sigma_u^2)$$

$$e_{ij} \sim N(0, \sigma_e^2). \tag{2}$$

As such, we can say that we are 'partially pooling' our data by assuming that our higher-level entities, though not identical, come from a single distribution σ_u^2, which is estimated from the data, much like the occasion-level variance σ_e^2, and can itself be interpreted substantively.

These models must not only be specified but also estimated on the basis of assumptions. Beck and Katz (2007) show that, with respect to TSCS data, RE models perform well, even when the Normality assumptions are violated[8]. As such they are preferred to both 'complete pooling' methods, which assume no differences between higher-level entities, and FE, which do not allow for the estimation of higher-level, time-invariant parameters or residuals (see sections 4 and 5). Shor et al. (2007) use similar methods, but estimated using Bayesian Markov Chain Monte Carlo (MCMC) (rather than Maximum likelihood) estimation,

[8] Outliers, however, are a different matter, but these can be dealt with using dummy variables for those outliers in a RE framework; see section 10.

which they find produces as good, or better[9], estimates to maximum likelihood RE and other methods.

3 The problem of omitted variable bias and endogeneity in Random Effects models

Considering this evidence, one must consider why it is that RE is not employed more widely, and remains rarely used in disciplines such as economics and political science. The answer lies in the exogeneity assumption of RE models: that the residuals are independent of the covariates; in particular the assumptions concerning the occasion-level covariates and the two variance terms, such that

$$E(u_j|x_{ij}, z_j) = E(e_j|x_{ij}, z_j) = 0.$$

In most practical applications this is synonymous with

$$\begin{aligned} Cov(x_{ij}, u_j) &= 0 \\ Cov(x_{ij}, e_{ij}) &= 0. \end{aligned} \tag{3}$$

The fact is that the above assumptions[10] often do not hold in many standard RE models as formulated in equation 1. Unfortunately, little attention has been paid to the substantive reasons why not. Indeed the discovery of such endogeneity has regularly led to the abandonment of RE in favour of FE estimation, which models out higher-level variance and

[9] The reason for this is that there is 'full error propagation' in Bayesian estimation as the uncertainty in both constituent parts of the model are taken into account, so that the variances of the random part are estimated on the basis that the fixed part are estimates and not known values, and vice versa. Simulations have shown that the improvement of MCMC estimated models over likelihood methods are greatest when there are there a small number of higher-level units, for example few countries (Browne and Draper, 2006, Stegmueller, 2013).

[10] An additional assumption implied here is that $Cov(z_j, u_{0j}) = 0$. Whilst this is an important assumption, it is not a good reason to choose FE as the latter cannot estimate the effect of z_j at all.

makes any correlations between that higher-level variance and covariates irrelevant, without considering the source of the endogeneity. This is unfortunate because the source of the endogeneity is often itself interesting and worthy of modelling explicitly.

This endogeneity most commonly arises as a result of multiple processes related to a given time-varying covariate[11]. In reality such covariates contain two parts: one that is specific to the higher-level entity which does not vary between occasions, and one which represents the difference between occasions, within higher-level entities:

$$x_{ij} = x_j^B + x_{ij}^W .$$

(4)

These two parts of the variable can have their own different effects: called 'between' and 'within' effects respectively, which together comprise the total effect of a given level 1, time-varying, variable. This division is inherent to the hierarchical structure present in both FE and RE models.

In equation 1 above, it is assumed that the within and the between effects are equal (Bartels, 2008). That is, a one-unit change in x_{ij} for a given higher-level entity has the same statistical effect (β_1) as being a higher-level entity with an inherent time-invariant value of x_{ij} that is 1 unit greater. Whilst this might well be the case, there are clearly many examples where this is unlikely. Considering an example of TSCS country data, an increase in equality may have a different effect to generally being an historically more equal country, for example due to some historical attribute(s) (such as colonialism) of that country.

[11] Whilst there may be other additional causes for correlation between x_{ij} and e_{ij}, this is the only cause of correlation between x_{ij} and u_j.

Indeed, as Snijders and Bosker (2012 p60) argue, "it is the rule rather than the exception that within-group regression coefficients differ from between-group coefficients."

Where the within and between effects are different, β_1 in equation 1 will be an uninterpretable weighted average of the two processes (Krishnakumar, 2006, Neuhaus and Kalbfleisch, 1998, Raudenbush and Bryk, 2002 p137) whilst variance estimates are also affected (Grilli and Rampichini, 2011). This can be thought of as omitted variable bias (Bafumi and Gelman, 2006, Palta and Seplaki, 2003); because the between effect is omitted, β_1 attempts to account for both the within and the between effect of the covariate on the response, and if the two effects are different, it will fail to account fully for either. The variance that is left unaccounted for will be absorbed into the error terms u_{0j} and e_{0ij}, which will consequently both be correlated with the covariate, violating the assumptions of the RE model. When viewed in these terms, it is clear that this is a substantive inadequacy in the theory behind the RE model, rather than just a statistical misspecification (Spanos, 2006) requiring a technical fix.

The word 'endogenous' has multiple forms, causes and meanings. It can be used to refer to bias caused by omitted variables, simultaneity, sample selection or measurement error (Kennedy, 2008 p139). These are all different problems that should be dealt with in different ways, and as such we consider the term misleading and, having explained it, do not use it in the rest of the article. The form of the problem that this article deals with is described rather more clearly by Li (2011) as 'heterogeneity bias', and we use that

terminology from now on. Our focus on this does not deny the existence of other forms of bias that cause and/or result from correlated covariates and residuals[12].

4 Fixed Effects Estimation

The rationale behind FE estimation is simple and persuasive, explaining why it is so regularly used in many disciplines. To avoid the problem of heterogeneity bias, all higher-level variance, and with it any between effects, are controlled out using the higher-level entities themselves (Allison, 2009), included in the model as dummy variables D_j:

$$y_{ij} = \sum_{j=1}^{j} \beta_{0j} D_j + \beta_1 x_{ij} + e_{ij}.$$

(5)

To avoid having to estimate a parameter for each higher-level unit, the mean for higher-level entity is taken away from both sides of equation 5, such that:

$$(y_{ij} - \bar{y}_j) = \beta_1 (x_{ij} - \bar{x}_j) + (e_{ij} - \bar{e}_j).$$

(6)

Because FE models only estimate within effects, they cannot suffer from heterogeneity bias. However, this comes at the cost of being unable to estimate the effects of higher-level processes, so RE is often preferred where the bias does not exist. In order to test for the existence of this form of bias in the standard RE model as specified in equation 1, the Hausman specification test (Hausman, 1978) is often used. This takes the form of a comparison between the parameter estimates of the FE and the RE model (Greene, 2012,

[12] Although we do deny that FE models are any better able to deal with these other forms of bias than RE models.

Wooldridge, 2002). This is done via a Wald test of the difference between the vector of coefficient estimates of FE and that of RE.

The Hausman test is regularly deployed as a test for whether RE can be used, or whether FE estimation should be used instead (for example Greene, 2012 p421). However, it is problematic when the test is viewed in terms of fixed and random effects, and not in terms of what is actually going on in the data. A negative result in a Hausman test tells us only that the between effect is not significantly biasing an estimate of the within effect in equation 1. It "is simply a diagnostic of one particular assumption behind the estimation procedure usually associated with the random effects model... it does not address the decision framework for a wider class of problems" (Fielding, 2004 p6). As we show later, the RE model which we propose in this paper solves the problem of heterogeneity bias described above and so makes the Hausman test, as a test of FE against RE, redundant. It is "neither necessary nor sufficient" (Clark and Linzer, 2012 p2) to use the Hausman test as the sole basis of a researcher's ultimate methodological decision.

5 Problems with Fixed Effects models

Clearly there are advantages to the FE model of equation 5-6 over the RE models in equation 1. By clearing out any higher-level processes, the model deals only with occasion-level processes. In the context of longitudinal data, this means considering differences over time, controlling out higher-level differences and processes absolutely and supposedly "getting rid of proper nouns" (King, 2001 p504), that is distinctive, specific characteristics of higher-level units. This is why it has become the "gold standard" method (Schurer and Yong, 2012 p1) in many disciplines. There is no need to worry about heterogeneity bias and β_1 can be thought to represent the 'causal effect'.

However, by removing the higher-level variance, FE models lose a large amount of important information. No inferences can be made about that higher-level variance, including whether or not that variance is significant (Schurer and Yong, 2012 p14). As such it is impossible to measure the effects of time-invariant variables at all, because all degrees of freedom at the higher level have been consumed. Where time-invariant variables are of particular interest this is obviously critical. And yet even in these situations, researchers have suggested the use of FE, on the basis of a Hausman test. For example, Greene's (2012 p420) textbook gives an example of a study of the effect of schooling on future wages:

> "The value of the [Hausman] test statistic is 2,636.08. The critical value from the chi-squared table is 16.919 so the null hypothesis of a random effects model is rejected. We conclude that the fixed effects model is the preferred specification for these data. This is an unfortunate turn of events, as the main object of the study is the impact of education, which is a time invariant variable in this sample."

Unfortunate indeed! To us, explicating a method which fails to answer your research question is nonsensical.

Furthermore, because the higher-level variance has been controlled out, any parameter estimates for time-varying variables deal with only a small subsection of the variance in that variable. Only within effects can be estimated, that is the lower level relationship net of any higher level attributes, and so nothing can be said about between effects or a general effect (if one exists) of a variable; studies which make statements about such effects on the basis of FE models are over-interpreting their results. Beck and Katz (Beck, 2001, Beck and Katz, 2001) consider the example of the effect of a rarely changing variable, democracy, on a binary variable representing whether a pair of countries are at peace or at war (Green et al., 2001, see also King, 2001, Oneal and Russett, 2001). They show that estimates obtained

under FE fail to show any relation between democracy and peace because it filters out all the effects of unchanging, time-invariant peace, which has an effect on time variant democracy. In other words, time-invariant processes can have effects on time-varying variables, which are lost in the FE model. Countries that do not change their political regime, or do not change their state of peace (that is most countries), are effectively removed from the sample. Whilst this problem applies particularly for rarely changing, almost time-invariant variables (Plümper and Troeger, 2007), any time-varying covariate can have such time-invariant 'between' effects, which can be different from time-varying effects of the same variable, and these processes cannot be assessed in a FE model. Only a RE model can allow these processes to be modelled simultaneously.

6 Plümper and Troeger's (2007) fixed effects vector decomposition

A method proposed by Plümper and Troeger (2007) allows time invariant variables to be modelled, within the framework of the FE model. They use a FE model before 'decomposing' the vector of fixed effects dummies into that explained by a given time-invariant (or rarely changing) variable, and that which is not. They begin by estimating a standard dummy variable fixed effects model as in equation 5:

$$y_{ij} = \sum_{j=1}^{j} \beta_{0j} D_j + \beta_1 x_{ij} + e_{ij}. \tag{7}$$

Here, D_j is a series of higher-level entity dummy variables, each with an associated intercept coefficient β_{0j}. Plümper and Troeger then regress in a separate higher-level model the vector of these estimated fixed effects coefficients on time-invariant variables, such that

$$\beta_{0j} = \beta_0 + \beta_2 z_j + R_j$$

(8)

where z_j is a (series of) higher level variable(s) and R_j is the residual. This equation can be rearranged so that, once estimated, the values of R_j can be estimated as

$$R_j = \beta_{0j} - \beta_2 x_{2j} - \beta_0.$$

(9)

Finally, equation 8 is substituted into equation 7 such that

$$y_{ij} = \sum_{j=1}^{j} (\beta_0 + \beta_2 z_j + R_j) D_j + \beta_1 x_{1ij} + e_{ij}$$
$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 z_j + \beta_3 R_j + e_{ij}.$$

(10)

where β_3 will equal exactly one (Greene, 2012 p405). The residual higher-level variance not explained by the higher-level variable(s) is modelled as a fixed effect leaving no higher-level variance unaccounted for. As such the model is very similar to a RE model (equation 1), which does a similar thing but in a single overall model[13]. Stage 1 (equation 7) is equivalent to the RE micro model, stage 2 (equation 8) to the macro model and stage 3 (equation 10) to the combined model. Just as with RE, the higher-level residual is assumed to be Normal (from the regression in equation 8). What it does do differently is also control out any

[13] In the early stages of the development of the multilevel model, a very similar process to the two-stage FEVD model was used to estimate processes at multiple levels (Burstein et al., 1978, Burstein and Miller, 1980), before being superseded by the modern multilevel, RE model in which an overall model is estimated (Raudenbush and Bryk, 1986). As Beck (2005 p458) argues: "perhaps at one time it could have been argued that one-step methods were conceptually more difficult, but, given current training, this can no longer be an excuse worth taking seriously."

between effect of x_{1ij} in the estimation of β_1, meaning these estimates will only include the within effect, as in standard FE models.

The Fixed effects vector decomposition (FEVD) estimator has been criticised by many in econometrics, who argue that the standard errors are likely to be incorrectly estimated (Breusch et al., 2011a, b, Greene, 2011a, b, 2012). Plümper and Troeger (2011) do provide a method for calculating more appropriate standard errors, and so the FEVD model does work (at least with balanced data – see section 8) when this method is utilised. However, our concern is that it retains many of the other flaws of FE models which we have outlined above. It remains much less generalisable than a RE model – it cannot be extended to three (or more) levels, nor can coefficients be allowed to vary (as in a random coefficients model – see section 8). It does not provide a nice measure of variance at the higher level, which is often interesting in its own right. Finally, it is heavily parameterised, with a dummy variable for each higher-level entity in the first stage, and so can be relatively slow to run when there are a large number of higher level units.

Plümper and Troeger also attempt to estimate the effects of 'rarely changing' variables, and their desire to do so by FE modelling suggests to us that they do not fully appreciate the difference between within and between effects. Whilst they do not quantify what rarely changing means, their motivation is in getting significant results where FE produces insignificant results. FE models only estimate within effects, and so an insignificant effect of a rarely changing variable should be taken as saying that there is no evidence for a within-effect of that variable. When Plümper and Troeger use FEVD to estimate the effects of rarely changing variables, they are in fact estimating between effects. Using FEVD to estimate the effects of rarely changing variables is not a technical fix for the high variance of

within effects in FE models – it is shifting the goalposts and measuring something different. Furthermore, if between effects of rarely changing variables are of interest, then there is no reason why the between effects of other time-varying variables would not be, and so these should potentially be modelled as time-invariant variables as well.

7 A Random Effects solution to heterogeneity bias

What is needed is a solution, within the parsimonious, flexible RE framework, which allows for heterogeneity bias not simply to be corrected, but for it to be explicitly modelled. As it turns out the solution is well documented, starting from a paper by Mundlak (1978a). By understanding that heterogeneity bias is the result of attempting to model two processes in one term (rather than simply a cause of bias to be corrected), Mundlak's formulation simply adds one additional term in the model for each time-varying covariate that accounting for the between effect: that is, the higher-level mean. This is treated in the same way as any higher-level variable. As such in the simple case the micro and macro models respectively are:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij}$$

and

$$\beta_{0j} = \beta_0 + \beta_2 z_j + \beta_3 \bar{x}_j + u_j.$$

This combines to form

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_3 \bar{x}_j + \beta_2 z_j + (u_j + e_{ij}) \quad (11)$$

where x_{ij} is a (series of) time variant variables, whilst $\bar{x}_j$ is the higher-level entity j's mean and as such the time-invariant component of those variables (Snijders and Bosker, 2012

p56). Here β_1 is an estimate of the within effect (as the between effect is controlled by $\bar{x}_j$); β_3 is the 'contextual' effect which explicitly models the difference between the within and the between effect. Alternatively, this can be rearranged by writing β_2 explicitly as this difference (Berlin et al., 1999):

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + (\beta_4 - \beta_1)\bar{x}_j + \beta_2 z_j + (u_j + e_{ij}).$$

This rearranges to:

$$y_{ij} = \beta_0 + \beta_1 (x_{ij} - \bar{x}_j) + \beta_4 \bar{x}_j + \beta_2 z_j + (u_j + e_{ij}). \quad (12)$$

Now β_1 is the within effect and β_4 is the between effect of x_{ij} (Bartels, 2008, Leyland, 2010). This 'within-between' formulation (see table 1) has three main advantages over Mundlak's original formulation. First, with temporal data it is more interpretable, as the within and between effects are clearly separated (Snijders and Bosker, 2012 p58). Second, in the first formulation, there is correlation between x_{ij} and $\bar{x}_j$; By group mean centring x_{ij}, this collinearity is lost, leading to more stable, precise estimates (Raudenbush, 1989). Finally, if multicollinearity exists between multiple $\bar{x}_j$s and other time-invariant variables, $\bar{x}_j$s can be removed without the risk of heterogeneity bias returning to the occasion-level variables (as in the within model in table 1)[14].

[Table 1 about here]

Just as before, the residuals at both levels are assumed to be Normally distributed:

[14] Instead of using the higher level unit mean (an aggregate variable), Clarke et al. (2010) suggest using global (Diez-Roux, 1998) unit characteristics that are correlated with that mean. These global variables express the causal mechanism underlying the association expressed by β_4, which may not be linear as is assumed by models 10 and 11. Including $\bar{x}_j$ would be over-controlling in this case, and such a model has a different interpretation of the higher-level residual, but it is harder to reliably control out all (or even most) of the between effect from the within effect without using $\bar{x}_j$ (Clarke, et al., 2010) in equation 11. However, this is not a problem when using the formulation in equation 12 as the within variable is already group mean centred, so the inclusion of $\bar{x}_j$ is optional depending on the research question at hand, as in the 'within' model in table 1.

$$u_j \sim N(0, \sigma_u^2)$$

$$e_{ij} \sim N(0, \sigma_e^2).$$

As can be seen, this approach is algebraically similar to the FEVD estimator (equation 10) – the mean term(s) are themselves interpretable time-invariant variables[15] (Begg and Parides, 2003), measuring the propensity of an higher-level entity to be x_{ij} (in the binary case) or the average level of x_{ij} (in the continuous case) across the sample time-period[16]. There are a few differences. First, estimates for the effects of time-invariant variables are controlled for by the means of the time-varying variables. Whilst this could be done in stage 2 of FEVD, it rarely is and nor is it suggested by Plümper and Troeger (2007) except for 'rarely changing' variables. Second, correct standard errors are automatically calculated, accounting for "multiple sources of clustering" (Raudenbush, 2009 p473). Crucially, there can be no correlation between the group mean centred covariate and the higher-level variance because the group mean centred covariate has a mean of 0 for each higher-level entity j. Equally, at the higher level the mean term is no longer constrained by level 1 effects, so is free to account for all the higher-level variance associated with that variable. As such, the estimate of β_1 in equations 11 and 12 above will be identical to that obtained by FE, as Mundlak (1978a p70) stated clearly:

[15] Because of this, the number of higher level units in the sample must be considered, and as such caution should be taken regarding how many higher level variables (including $\bar{x}_j$s) the model can estimate reliably. The MLPowSim software (Browne et al., 2009) can be used to judge this in the research design phase.

[16] Note that when interpreting these terms, we are usually interested in general, latent characteristics of an individual which are invariant beyond the sample period. From this perspective it is not the case that we are conditioning on the future (as argued by Kravdal, 2011), any more than with any other time invariant variable. However because these means are measured from a finite sample, they are subject to measurement error and their coefficients subject to bias. This can be corrected for by shrinking them back towards the grand mean, in a similar way to the residuals, through equation 13 (see Grilli and Rampichini, 2011, Shin and Raudenbush, 2010). However more detailed explication of this is beyond the scope of this paper.

> "when the model is properly specified, the GLSE [that is RE] is identical to the "within" [that is, the FE] estimator. Thus there is only one estimator. The whole literature which has been based on an imaginary difference between the two estimators ... is based on an incorrect specification which ignores the correlation between the effects and the explanatory variables."

Whilst it is still possible that there is correlation between the group mean centered x_{ij} and e_{ij}, and between $\bar{x}_j$ (and other higher-level variables) and u_j (Kravdal, 2011), this is no more likely than in FE models for the former and aggregate regression for the latter because we have accounted for the key source of this correlation by specifying the model correctly (Bartels, 2008).[17] After all, "all models are wrong; the practical question is how wrong do they have to be to not be useful" (Box and Draper, 1987 p74). How useful the model is depends, as with any model, on how well the researcher has accounted for possible omitted variables, simultaneity, or other potential model misspecifications.

We see the FE model as a constrained form of the RE model[18], meaning that the latter can encompass the former but not vice-versa. By using the random effects configuration, we keep all the advantages associated with RE modelling[19]. First, the 'problem' of heterogeneity bias across levels is not simply solved; it is explicitly modelled. The effect of x_{ij} is separated into two associations, one at each level, which are interesting, interpretable, and relevant to the researcher (Enders and Tofighi, 2007 p130). Second, by

[17] If covariates remain correlated with residuals (for example as a result of simultaneity, or other omitted variables), they can potentially be dealt with within this RE framework through other means, such as instrumental variable methods (Heckman and Vytlacil, 1998) using simultaneous equations (Steele et al., 2007), assuming of course that appropriate instruments can be found. Whilst all heterogeneity bias of lower-level variables has been dealt with, a variant of the Hausman-Taylor IV estimator (Greene, 2012 p434, Hausman and Taylor, 1981) can be used to deal with correlated time-invariant variables (Chatelain and Ralf, 2010).

[18] Demidenko (2004 p.54-55) proves that the FE model is equivalent to a RE model in which the higher level variance is constrained to be infinite.

[19] Note that it is still necessary to use RE estimation methods (rather than OLS) in order for correct SEs to be calculated.

assuming Normality of the higher-level variance, the model need only estimate a single term for each level (the variance), which are themselves useful measures, allowing calculation of the variance partitioning coefficient (VPC)[20], for example. Further, higher-level residuals (conditional on the variables in the fixed part of the model) are precision-weighted or shrunken by multiplying by the higher-level entity's reliability λ_j (see Snijders and Bosker, 2012 p62),[21] calculated as:

$$\lambda_j = \frac{\sigma_u^2}{\sigma_u^2 + (\sigma_e^2 / n_j)} \tag{13}$$

where n_j is the sample size of higher-level entity j, σ_u^2 is the between-entity variance, and σ_e^2 is the variance within higher-level entities, between occasions. One can thus estimate reliable residuals for each higher-level entity that are less prone to measurement error than FE dummy coefficients. By partially pooling through assuming that u_j comes from a common distribution with a variance that has been estimated from the data, we can obtain much more reliable predictions for individual higher-level units (see Rubin, 1980, for an early example of this)[22]. Whilst this is rarely of interest in individual panel data, it is likely to be of interest with TSCS data with repeated measures of countries, as we see in section 10.

The methods which we are proposing here are beginning to be taken up by researchers, under the guise of a 'hybrid' or 'compromise' approach between FE and RE (Allison, 2009 p23, Bartels, 2008, Greene, 2012 p421). This is to misrepresent the nature of the model.

[20] The VPC is the proportion of variance that occurs at level 2. In the simple 2-level RE case it is calculated as $\frac{\sigma_u^2}{\sigma_u^2+\sigma_e^2}$, and is a standardised measure of the similarity between higher level units.

[21] A detailed comparison between the fixed and random effects estimates is given algebraically and empirically in Jones and Bullen (1994)

[22] We are assuming here that higher level units come from a single distribution. This is usually a reasonable assumption, and it can be readily evaluated, as we see in the example in section 10.

There is nothing Fixed-Effects-like about the model at all – it is a RE model with additional time-invariant predictors. Perhaps as a consequence of this potentially misleading terminology, many of those who use such models fail to recognise its potential as a RE model. Allison (2009 p25), for example, argues that the effects of the mean variables ($\bar{x}_j$) "are not particularly enlightening in themselves", whilst many have suggested using the formulation as a form of the Hausman test and use the results to choose between fixed and random effects (Allison, 2009 p25, Baltagi, 2005, Greene, 2012 p421, Hsiao, 2003 p50, Wooldridge, 2002 p290, 2009). Thus, β_3 in equation 11 is thought of simply as a measure of 'correlation' between x_{ij} and u_j and when $\beta_3 \neq 0$ in equation 11, or $\beta_1 \neq \beta_4$ in equation 12, the Hausman test fails and it is argued the FE model should be used. It is clear to us (and to Skrondal and Rabe-Hesketh, 2004 p53, Snijders and Berkhof, 2007 p145), however, that the use of this model makes that choice utterly unnecessary.

To reiterate: the Hausman test is not a test of FE versus RE; it is a test of the similarity of within and between effects. A RE model that properly specifies the within and between effects will provide identical results to FE, regardless of the result of a Hausman test. Furthermore, between effects, other higher-level variables and higher level residuals, none of which can be estimated with FE, should not be dismissed lightly; they are often enlightening, especially for meaningful entities such as countries. For these reasons, and the ease with which they can now be fitted in most statistical software packages, RE models are the obvious choice.

8 Simulations

We now present simulations results which show that, under a range of situations, the RE solution that we propose performs at least as well as the alternatives on offer – it predicts the same effects as both FE and FEVD for time varying variables, and the same results for time invariant variables as FEVD. Furthermore, the simulations show that standard errors are poorly estimated by FEVD when there is imbalance in the data.

The simulations are similar to those conducted by Plümper and Troeger (2007), using the following underlying DGP:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_{3C} \bar{x}_{3j} + \beta_4 z_{1j} + \beta_5 z_{2j} + \beta_6 z_{3j} + u_j + e_{ij} \quad (14)$$

where

$$\beta_0 = 1, \beta_1 = 0.5, \beta_2 = 2, \beta_3 = -1.5, \beta_4 = -2.5, \beta_5 = 1.8\, \beta_6 = 3.$$

In order to simulate correlation between x_{3ij} and u_j, the value of β_{3C} varies (-1, 0, 1, 2) between simulations. This parameter is also estimated in its own right – as we have argued, it is often of substantive interest in itself. We also vary the extent of correlation between z_{3j} and u_j (-0.2, 0, 0.2, 0.4, 0.6). All variables were generated to be Normally distributed with a mean of zero – fixed part variables with a standard deviation of 1, level 1 and 2 residuals with standard deviations of 3 and 4 respectively. In addition, we varied the sample size - both the number of level 2 units (100, 30) and the number of time points (20, 70). Additionally we tested the effect of imbalance in the data (no missingness, 50% missingness in all but five of the higher level units) on the performance of the various estimators. The simulations were run in Stata using the xtreg and xtfevd commands.

For each simulation scenario, the data were generated and models estimated 1000 times, and three quantities were calculated: bias, root mean square error (RMSE) and optimism, calculated as in Shor et al (2007) and in line with the simulations presented by Plümper and Troeger (2007, 2011). Bias is the mean of the ratios of the true parameter value to the estimated parameter, and so a value of 1 suggests that the model estimates are on average exactly correct. RMSE also assesses bias, as well as efficiency, where the lower the value, the more accurate and precise the estimator. Finally optimism evaluates how the standard errors compare to the true sampling variability of the simulations; values greater than 1 suggest that the estimator is overconfident in its estimates, whilst values below one suggest that they are more conservative than necessary.

Table 2 presents the results from some permutations of the simulations when the data is balanced. As can be seen, and as expected, the standard RE estimator is outperformed by the other estimators, because of bias resulting from the omission of the between effect associated with X3 from the model. It can also be seen that the within-between RE model (REWB) performs at least as well as both FE and FEVD for all three measures. What is more surprising is the effect of data imbalance on the performance of the estimators – whilst for RE, FE and REWB the results remain much the same, the standard errors are estimated poorly by the FEVD – too high (type 2 errors) for lower level variables and too low (type 1 errors) for higher level variables. The online appendix shows that this result is repeated for all the simulation scenarios that we tested, regardless of the size of correlations present in the data and the data sample size. It is clear that it would be unwise to use the FEVD with unbalanced data, and even when data is balanced, Mundlak's (1978a) claim, that the models will produce identical results, is justified.

[Table 2 and 3 about here]

Having shown that the within-between random effects model produces results which are at least as unbiased as alternatives including FE and FEVD, the question remains why one should choose the random effects option over these others. If higher level variables and/or shrunken residuals are not of substantive interest, Why not simply estimate a FE regression (or the FEVD estimator if time-invariant variables or other between effects happen to be of interest and the data is balanced)? The answer is two-fold. First, with the ability to estimate both effects in a single model (rather than the three steps of the FEVD estimator), the RE model is more general than the other models. We believe it is valuable to be able to model things in a single coherent framework. Second, and more importantly, the RE model can be extended to allow for variation in effects across space and time to be explicitly modelled, as we show in the following section. That is, whilst FE models assume a priori that there is a single effect that affects all higher-level units in the same way, the RE framework allows for that assumption to be explicitly tested. As the example in section 10 shows, this does not simply provide additional results to those already found - failing to do this can lead to results that are seriously and substantively misleading.

9 Extending the basic model: Random Coefficient Models and cross-level interactions

We have argued that the main advantage of RE models is their generalisability and extendibility, and this section outlines one[23] such extension: the random coefficient model (RCM). This allows the effects of β coefficients to vary by the higher-level entities (Bartels,

[23] Other potential model extensions could include 3-level models, or multiple membership or cross-classified (Raudenbush 2009) data structures.

2008, Mundlak, 1978b, Schurer and Yong, 2012). Heteroscedasticity at the occasion level can also be explicitly modelled by including additional random effects at level 1. As such, our model could become

$$y_{ij} = \beta_{0j} + \beta_{1j}(x_{ij} - \bar{x}_j) + e_{0ij} + e_{1ij}(x_{ij} - \bar{x}_j)$$

where

$$\beta_{0j} = \beta_0 + \beta_4\bar{x}_j + \beta_2 z_j + u_{0j}$$
$$\beta_{1j} = \beta_1 + u_{1j}.$$

These equations (one micro and two macro) combine to form:

$$y_{ij} = \beta_0 + \beta_1(x_{ij} - \bar{x}_j) + \beta_4\bar{x}_j + \beta_2 z_j + [u_{0j} + u_{1j}(x_{ij} - \bar{x}_j) + e_{0ij} + e_{1ij}(x_{ij} - \bar{x}_j)] \quad (15)$$

with the following distributional assumptions:

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_{u0}^2 & \\ \sigma_{u0u1} & \sigma_{u1}^2 \end{bmatrix}\right)$$

$$\begin{bmatrix} e_{0ij} \\ e_{1ij} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_{e0}^2 & \\ \sigma_{e0e1} & \sigma_{e1}^2 \end{bmatrix}\right).$$

These variances and covariances can be used to form quadratic 'variance functions' (Goldstein, 2010 p73) to see how the variance varies with $(x_{ij} - \bar{x}_j)$. At the higher level, the total variance is calculated by

$$var[u_{0ij} + u_{1ij}(x_{ij} - \bar{x}_j)] = \sigma_{u0}^2 + 2\sigma_{u0u1}(x_{ij} - \bar{x}_j) + \sigma_{u1}^2(x_{ij} - \bar{x}_j)^2 \quad (16)$$

and at level 1, it is

$$var[e_{0ij} + e_{1ij}(x_{ij} - \bar{x}_j)] = \sigma_{e0}^2 + 2\sigma_{e0e1}(x_{ij} - \bar{x}_j) + \sigma_{e1}^2(x_{ij} - \bar{x}_j)^2. \quad (17)$$

These can often be substantively interesting, as well as being a correction for misspecification of a model that would otherwise assume homogeneity at each level (Rasbash et al., 2009 p106). As such, even when time-invariant variables are not of interest, the RE model is preferable because it means that "a richer class of models can be estimated" (Raudenbush, 2009 p481), and rigid assumptions of FE and FEVD can be relaxed.

RCMs additionally allow cross-level interactions between higher- and lower-level variables. In the TSCS case, that is an interaction between a variable measured at the country level and one measured at the occasion level. This is achieved by extending equation 15 to, for example:

$$y_{ij} = \beta_{0j} + \beta_{1j}(x_{ij} - \bar{x}_j) + e_{0ij} + e_{1ij}(x_{ij} - \bar{x}_j)$$

$$\beta_{0j} = \beta_0 + \beta_4\bar{x}_j + \beta_2 z_j + u_{0j}$$

$$\beta_{1j} = \beta_1 + \beta_5\bar{x}_j + u_{1j}$$

which combine to form:

$$y_{ij} = \beta_0 + \beta_1(x_{ij} - \bar{x}_j) + \beta_4\bar{x}_j + \beta_5(x_{ij} - \bar{x}_j)\bar{x}_j + \beta_2 z_j + [u_{0j} + u_{1j}(x_{ij} - \bar{x}_j) + e_{0ij} + e_{1ij}(x_{ij} - \bar{x}_j)]. \quad (18)$$

The models can thus give an indication of whether the effect of a time-varying predictor varies by time-invariant predictors (or vice-versa), and this is quantified by the coefficient β_5. Note that these could include interactions between the time variant and time-invariant parts of the same variable, as is the case above, or could involve other time-invariant variables. The possibility of such interactions is not new (Davis et al., 1961) and have been an established part of the multilevel modelling literature for many years (Jones and Duncan,

1995 p33). Whilst the interaction terms themselves can be included in a FE model (for example see Boyce and Wood, 2011, Wooldridge, 2009), it is only when they are considered together with the additive effects of the higher-level variable (β_4) that their full meaning can be properly established. This can only be done in a random coefficient model. Such relationships ought to be of interest to any researcher studying time-varying variables. If the effect of a time-varying education policy is different for boys and girls, the researcher needs to know this. It is even conceivable that such relationships could be in opposite directions for different types of higher-level entity. In which case, a FE study that suggests a policy generally helps everyone could be hiding the fact that it actually hinders certain types of people. Resources could be wasted applying a policy to individuals that are harmed by it. Following Pawson (2006), we believe that context should be central to any evidence-based policy.

To reiterate this point: even when time-invariant variables are not directly relevant to the research question itself, it is important to think about what is happening at the higher level, in a multilevel RE framework. Simpler models that control out context assume that occasion-level covariates have only 'stylised' (see Clark, 1998, Kaldor, 1961, Solow, 1988 p2) mean effects that affect all higher-level entities in exactly the same way. This leads to nice simple conclusions (a policy either works or does not), but it misses out important information about what is going on:

> "Continuing to do individual-level analyses stripped out of its context will never inform us about how context may or may not shape individual and ecological outcomes."
>
> (Subramanian et al., 2009a p355)

The example below shows the advantages in terms of substantive insights that can be gained from the RE framework, and the dangers of failing to implement the extensions suggested in the present section. A Hausman test would suggest that, for this dataset, a FE model should be used; we show that doing so leads to considerably impoverished results.

10 Example: the effect of democracy on trade liberalism

Our example uses TSCS data to look at the effect of democracy on trade openness in developing countries. The data consists of a measure of a country's statutory tariff rate as the dependent variable (with low tariffs reflecting trade openness), and independent variables including a polity score (measured between -10 and +10, where high values indicate greater democracy), GDP per capita, the natural log of the country population, and the year of measurement[24], measured on occasions (level 1) between 1980 and 1999 for 101 countries (level 2)[25] (see Table 4). Milner and Kubota (2005) use FE estimation (see model 2 in table 5) and argue that their findings show that "more democratic regimes tend to have lower tariff rates" (p126). Of course this is an over-interpretation, as their FE model can only measures within-country effects – their results only actually suggest that a country *becoming* more democratic leads to lower tariff rates. Here, we reanalyse Milner and Kubota's data under a RE framework and show that even that conclusion is subject to considerable doubt.

[Table 4 about here]

[24] Note that Milner and Kubota also have more complex models with more control variables. Here we use their most simple model (model 1 in their table 2, p127) to illustrate our methodological argument as clearly as possible. We intend to make a more definitive critique in a later paper.

[25] Milner and Kubota's article suggested that their data ran from 1970. In fact, for all countries the data was subject to missingness until 1980. Note that the rest of the data is also subject to missingness or imbalance. However, more appropriate methods for dealing with missing data (Carpenter et al., 2011) are beyond the scope of this paper, so here we use listwise deletion on all cases with missing values in the predictors and outcome that we use.

Our models are as follows (see Tables 5 and 6)[26]:

1. A null RE model with no predictors (a simplification of equation 1)
2. A FE model, similar to that used in Milner and Kubota (2005) (equation 6)
3. A standard RE model which takes no account of heterogeneity bias (equation 1)
4. RE model with the within-between specification[27] (equation 12)
5. As 4 but with outlying intercepts included as a single dummy variable
6. As 5 but with the coefficient for within polity score allowed to vary at both level 1 and 2 (equation 15)
7. As 6 but with an outlier polity effect included as a differential slope in the fixed part
8. As 7 with a cross level interaction between the within and between components of democracy included (equation 18).

These models were fitted using MLwiN version 2.27[28] (Rasbash et al., 2013) with RIGLS restricted maximum likelihood estimation.[29]

Looking at Model 1, we can calculate the VPC from the two variance terms, and see that 58% of the variance in the response occurs at the higher-level, between individuals. As FE models only look at the occasion-level variance, they therefore can only consider 42% of the interesting variation that is going on in the dependent variable. Context, in this case

[26] Milner and Kubota additionally use an AR1 correction to allow autocorrelated residuals. We ran the simpler of our models with autocorrelated residuals and found that it did not affect our substantive conclusions. In order to keep this model as simple as possible for illustrative purposes, we therefore do not report the results with auto-correlated residuals.

[27] Note that the 'between' country means were calculated using the full data, prior to listwise deletion. The within components were calculated using the country means of the cut down data, to preserve orthogonality.

[28] These models can also be easily estimated in most major statistical software packages, including Stata, R and SAS. Code to implement the models in Stata using the 'runmlwin' command (Leckie and Charlton, 2013) can be found in the appendix.

[29] With the exception of the FE model, which was estimated using the xtreg command in Stata. MCMC results were largely the same, as would be expected due to the large number of higher level units.

individual difference, is being controlled out when it is at this higher level that most of the variance lies, meaning the majority of the variation in the data is effectively being ignored.

Comparing model 2 and 3, we see that the mis-specified standard RE model without level 2 means suffers from bias, particularly in the population variable (lnpop), the effect of which is vastly underestimated. However, the FE results (model 2) are identical to the within part of the RE estimates of model 4 which models the cause of this bias (different within and between effects) explicitly. In addition, including the mean term of polity in model 4 shows us that, in fact, there is no evidence for an effect of a country's average level of democracy over the period of measurement on free trade. Milner and Kubota's (2005, p126) conclusion that "more democratic regimes tend to have lower tariff rates" is in fact unsupported by this analysis.

[Tables 5 and 6 about here]

One of the characteristics of TSCS data is an interest in how individual entities, in this case countries, operate differently from each other. Milner and Kubota express this interest early in their article, drawing attention to specific countries that have experienced democratisation and trade liberalisation simultaneously. However, their method is unable to consider the heterogeneity of individual countries because their FE analysis controls out all country effects. In contrast the shrunken higher-level residuals in a RE model can be estimated to consider variation between countries. Figure 1 (obtained from model 4) shows this – there are three clear South Asian outliers with much higher differential intercepts than other countries. These cannot be thought of as part of the overall distribution of countries, so they are 'dummied out' as a set in model 5 to have their own differential

intercept and preserve the assumption of Normal residuals for those countries that remain in the random part as part of a common distribution[30].

[Fig 1 about here]

In Model 6 the coefficient associated with the within polity score is allowed to vary, and the associated random coefficient shrunken residuals are plotted in Figure 2. Again, we find that Bangladesh is a substantial outlier in its effect, having a much steeper negative slope than other countries; it is much more difficult to find these outliers with FE models. We included an interaction between this country's dummy and the within democracy variable (model 7) to allow it to have its own differential slope and remove it from the common distribution of higher-level effects. This caused the overall 'within' polity effect to become insignificant. The overall mean effect found by Milner and Kubota appears to be solely the result of a single outlying country, and this is made clear by figure 3. Their use of FE to get "rid of proper nouns" (King, 2001 p504) misleads because it is a specific entity (Bangladesh) rather than a common global effect that is driving the supposedly causal relationship. This shows the importance of assessing outliers in the effect as well as the constant, and this is difficult to do in a FE framework. In contrast, RCMs do this almost automatically, and by using dummy variables to model these outliers "the specifics of people and places are retained in a model, which still has a capacity for generalisation" (Jones, 2005 p255). Whilst India would be an even more extreme outlier in terms of its raw slope residual, it has very little variation in its within polity score, making its unusual slope much less reliable than that of Bangladesh. There is thus substantial shrinkage for India's slope residual (see equation 13), and the result is shown in figures 2 and 3, reflecting the fact that its effect on the mean

[30] The dummied variables are now effectively fixed effects. As they are no longer shrunk based on a common variance (see equation 13), the value of the dummy coefficient is greater than the values of the points plotted in figure 1.

coefficient is minimal in comparison to that of Bangladesh. FE dummy coefficients are un-shrunken so it is not possible to consider distinguishing between a reliably unusual country-specific effect, and an unreliably unusual one, in this way.

[Fig 2 and 3 about here]

A further advantage of random coefficient models is the potential to use variance functions to ascertain how variance changes with polity score (see Figure 4). We see that there is much greater variation (conditional on the fixed part of the model) between countries with a low within polity score than those with a high within polity score; assuming a general trend towards democracy over time, this suggests that countries tariff rates become more alike as they move towards democracy. At level 1, there was evidence of a linear variance function[31], whereby countries are slightly more volatile between occasions in their trade policy where there has been a move towards democracy.

[Fig 4 about here]

In model 8, a cross level interaction was included to attempt to explain the variation in the slopes with the within polity score seen in Figures 2 and 3. Whilst the overall effect of within-country democracy was and still is insignificant, there does appear to be differential effects for different countries. In fact Figure 5 shows that, for countries that are generally (historically, over the long term) undemocratic, the effect of an increase in democracy is in the opposite direction to that suggested by Milner and Kubota – as they become more democratic they tend to increase tariff rates. This is an interesting result, which suggests very different causal explanations to the uniform effect posited by Milner and Kubota. The

[31] Including the term σ^2_{e1} associated with within polity did not reduce the deviance, so there was no evidence for a full quadratic variance function. The linear variance function equation reduces from equation 17 to: $var[e_{0ij} + e_{1ij}(x_{ij} - \bar{x}_j)] = \sigma^2_{e0} + 2\sigma_{e0e1}(x_{ij} - \bar{x}_j)$. See Bullen et al. (1997).

world is messier and more heterogeneous than a FE model allows it to be, and that messiness needs to be considered before researchers can be sure of the substantive meaning of their results.

[Fig 5 about here]

11 Conclusions

In the introduction to his book on fixed effects models, Allison (2009 p2) criticises an early proponent of RE:

> "such characterisations are very unhelpful in a nonexperimental setting, however, because they suggest that a random effects approach is nearly always preferable. Nothing could be further from the truth."

We have argued in this paper that, in fact, the RE approach is nearly always preferable. We have shown that the main criticism of RE, the correlation between covariates and residuals, is readily solvable using the within-between formulation espoused here, although the solution is used all too rarely in RE modelling. This is why, in fact, Allison argues in favour of the same RE formulation that we have used, even though he calls it a 'hybrid' solution. Our strong position is not simply based on finding a technical fix, however. We believe that understanding the role of context, be it households, individuals, neighbourhoods, countries or whatever defines the higher level, is usually of profound importance to a given research question – one must model it explicitly, and that requires the use of a RE model that analyses and separates both the within and between components of an effect explicitly, and assesses how those effects vary over time and space rather than assuming heterogeneity away with FE:

> "heterogeneity is not a technical problem calling for an econometric solution but a reflection of the fact that we have not started on our proper business, which is trying to understand what is going on."
>
> (Deaton, 2010 p430)

This point is as much philosophical as it is statistical (Jones, 2010). We as researchers are aiming to understand the world. FE models attempt to do this by cutting out much of 'what is going on', leaving only a supposedly universal effect and controlling out differences at the higher level. In contrast, a RE approach explicitly models this difference, leading "to a richer description of the relationship under scrutiny" (Subramanian et al., 2009b p373). To be absolutely clear, this is not to say that within-between RE models are perfect – no model is. If there are only a very small number of higher level units, RE may not be appropriate. As with any model it is important to consider whether important variables have been omitted and whether causal interpretations are justified, using theory, particularly regarding time-invariant variables. No statistical model can act as a substitute for intelligent research design and forethought regarding the substantive meaning of parameters. However the advantages of within-between RE over the more restrictive FE are at odds with the dominance of FE as the 'default' option in a number of social science disciplines. We hope this article will go some way towards ending that dominance and stimulating much needed debate on this issue.

12 Appendix A: Stata code for the models

RE models can be fitted easily in Stata using the xtmixed command. However, for the most complex models (for example with complex variance at the occasion level), the command 'runmlwin' (Leckie and Charlton, 2013) can be used. This requires MLwiN to be installed on the computer; Stata specifies the model, runs it in MLwiN and transfers the results back to Stata. Below is the code for the models in tables 5 and 6, including the generation of within and between variables and interaction variables:

```
*to install runmlwin

 ssc install runmlwin, replace
global MLwiN_path "[pathway to MLwiN program for your computer]"



*load the dataset

use Milner1, clear

*generate mean variables

egen gdppc_mean = mean(gdppc), by(ctylabel)
egen polity_mean = mean(polity), by(ctylabel)
egen lnpop_mean = mean(lnpop), by(ctylabel)

*remove missing values

drop if missing(tariff)
drop if missing(gdppc)
drop if missing(polity)
drop if missing(lnpop)

*generate within variables

egen date_mean_new = mean(date), by(ctylabel)
egen gdppc_mean_new = mean(gdppc), by(ctylabel)
egen polity_mean_new = mean(polity), by(ctylabel)
egen lnpop_mean_new = mean(lnpop), by(ctylabel)

gen datew = date - date_mean_new
gen gdppcw = gdppc - gdppc_mean_new
gen polityw = polity - polity_mean_new
gen lnpopw = lnpop - lnpop_mean_new

drop gdppc_mean_new
drop polity_mean_new
```

```
drop lnpop_mean_new

*center variables

sum gdppc, meanonly
gen cgdppc = gdppc - r(mean)
sum date, meanonly
gen cdate = date - r(mean)
sum lnpop, meanonly
gen clnpop = lnpop - r(mean)

sum gdppc_mean, meanonly
gen cgdppc_mean = gdppc_mean - r(mean)
sum lnpop_mean, meanonly
gen clnpop_mean = lnpop_mean - r(mean)

*generate cross-level interactions etc

generate politywxpolity_mean = polityw*polity_mean

generate SAsia = 0
replace SAsia = 1 if ctylabel == 68
replace SAsia = 1 if ctylabel == 61
replace SAsia = 1 if ctylabel == 76

generate bangla = 0
replace bangla = 1 if ctylabel == 61

generate BanglaXPolityw = bangla*polityw

*generate the matrix for the linear variance function at level 1 in models 6-8

matrix A = (1,1,0)

*set the nature of the data (needed for xtreg)

tsset ctylabel date, yearly

*run the models

runmlwin tariff cons, level2(ctylabel: cons) ///
level1(date: cons) nopause rigls
estimates store REnull

xtreg tariff polity clnpop cgdppc cdate, fe
estimates store FE

runmlwin tariff cons polity clnpop cgdppc cdate, ///
level2(ctylabel: cons) level1(date: cons) nopause rigls
estimates store RE
```

```
runmlwin tariff cons politywlnpopw gdppcw datew polity_mean ///
clnpop_mean cgdppc_mean, level2(ctylabel: cons) ///
level1(date: cons) nopause rigls
estimates store REwb

runmlwin tariff cons politywlnpopw gdppcw datew polity_mean ///
clnpop_mean cgdppc_mean SAsia, ///
level2(ctylabel: cons) level1(date: cons) ///
initsprevious nopause rigls
estimates store mod5

runmlwin tariff cons politywlnpopw gdppcw datew polity_mean ///
clnpop_mean cgdppc_mean SAsia, ///
level2(ctylabel: cons politywlevel1(date: cons politywlelements(A)) ///
initsprevious nopause rigls
estimates store mod6

runmlwin tariff cons politywlnpopw gdppcw datew polity_mean ///
clnpop_mean cgdppc_mean SAsia BanglaXPolitywl///
level2(ctylabel: cons politywlevel1(date: cons politywlelements(A)) ///
initsprevious nopause rigls
estimates store mod7

runmlwin tariff cons politywlnpopw gdppcw datew polity_mean ///
clnpop_mean cgdppc_mean SAsia BanglaXPolitywpolitywxpolity_mean, ///
level2(ctylabel: cons politywlevel1(date: cons politywlelements(A)) ///
initsprevious nopause rigls
estimates store mod8
estimates table REnull FE RE REwb, se stats(deviance)

estimates table mod5 mod6 mod7 mod8, se stats(deviance)
```

13 References

Allison, Paul D. 2009. *Fixed effects regression models.* London: Sage.

Arceneaux, Kevin, and David W. Nickerson. 2009. Modeling certainty with clustered data: a comparison of methods. *Political Analysis,* 17 (2): 177-90.

Bafumi, Joseph, and Andrew Gelman. 2006. Fitting multilevel models when predictors and group effects correlate. *Annual Meeting of the Midwest Political Science Association*. Chicago, IL. Available at http://www.stat.columbia.edu/~gelman/research/unpublished/Bafumi_Gelman_Midwest06.pdf [Accessed 21st March 2012].

Baltagi, Badi H. 2005. *Econometric analysis of panel data.* 3rd edition. Chichester: Wiley.

Bartels, Brandon L. 2008. Beyond "fixed versus random effects": a framework for improving substantive and statistical analysis of panel, time-series cross-sectional, and multilevel data. *Political Methodology Conference*. Ann Arbor, MI. Available at http://home.gwu.edu/~bartels/cluster.pdf [Accessed 1 March 2012].

Beck, Nathaniel. 2001. Time-series-cross-section-data: What have we learned in the past few years? *Annual Review of Political Science,* 4: 271-93.

———. 2005. Multilevel analyses of comparative data: A comment. *Political Analysis,* 13 (4): 457-58.

———. 2007. From statistical nuisances to serious modeling: Changing how we think about the analysis of time-series-cross-section data. *Political Analysis,* 15 (2): 97-100.

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.

———. 2001. Throwing out the baby with the bath water: A comment on Green, Kim, and Yoon. *International Organization,* 55 (2): 487-95.

———. 2007. Random coefficient models for time-series-cross-section data: Monte Carlo experiments. *Political Analysis,* 15 (2): 182-95.

Begg, Melissa D, and Michael K Parides. 2003. Separation of individaul-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine,* 22: 2591-602.

Berlin, Jesse A., Stephen E. Kimmel, Thomas R. Ten Have, and Mary D. Sammel. 1999. An empirical comparison of several clustered data approaches under confounding due to cluster effects in the analysis of complications of coronary angioplasty. *Biometrics,* 55 (2): 470-76.

Box, George E, and Norman R Draper. 1987. *Empirical model-building and response surfaces.* USA: Wiley.

Boyce, Christopher J., and Alex M. Wood. 2011. Personality and the marginal utility of income: Personality interacts with increases in household income to determine life satisfaction. *Journal of Economic Behavior & Organization,* 78 (1-2): 183-91.

Breusch, Trevor, Mickael B. Ward, Hoa T M Nguyen, and Tom Kompas. 2011a. On the Fixed-Effects Vector Decomposition. *Political Analysis,* 19 (2): 123-34.

———. 2011b. FEVD: Just IV or just mistaken? *Political Analysis,* 19 (2): 165-69.

Browne, William J, Mousa G Lahi, and Richard MA Parker. 2009. A guide to sample size calculations for random effect models via simulation and the MLPowSim software package. University of Bristol. Available at http://www.bristol.ac.uk/cmm/software/mlpowsim/ [Accessed 21st June 2012].

Browne, William J., and David Draper. 2006. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis,* 1 (3): 473-513.

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.

Burstein, Leigh, Robert L Linn, and Frank J Capell. 1978. Analyzing multlevel data in the presence of heterogeneous within-class regressions. *Journal of educational statistics,* 3 (4): 347-83.

Burstein, Leigh, and Michael David Miller. 1980. Regression-based analyses if multilevel education data. *New directions for methodology of social and behavioral sciences,* 6: 194-211.

Carpenter, James R., Harvey Goldstein, and Michael G. Kenward. 2011. REALCOM-IMPUTE: Software for multilevel multiple imputation with mixed response types. *Journal of Statistical Software,* 45 (5): 1-14.

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*. Available at http://hal-paris1.archives-ouvertes.fr/docs/00/49/20/39/PDF/Chatelain_Ralf_Time_Invariant_Panel.pdf [Accessed 16th April 2012].

Clark, Gordon L. 1998. Stylized facts and close dialogue: Methodology in economic geography. *Annals of the Association of American Geographers,* 88 (1): 73-87.

Clark, Tom S, and Drew A Linzer. 2012. Should I use fixed or random effects? : Emory Universrity. Available at http://polmeth.wustl.edu/mediaDetail.php?docId=1315 [Accessed 3rd May 2012].

Clarke, Paul, Claire Crawford, Fiona Steele, and Anna Vignoles. 2010. The choice between fixed and random effects models: some considerations for educational research. *CMPO working paper*: University of Bristol. Available at http://www.bristol.ac.uk/cmpo/publications/papers/2010/wp240.pdf [Accessed 1st March 2012].

Davis, James A., Joe L. Spaeth, and Carolyn Huson. 1961. A technique for analyzing the effects of group composition. *American Sociological Review,* 26 (2): 215-25.

Deaton, Angus. 2010. Instruments, randomization, and learning about development. *Journal of Economic Literature,* 48 (2): 424-55.

Demidenko, Eugene. 2004. *Mixed models: theory and applications.* Hoboken, NJ: Wiley.

Diez-Roux, Ana V. 1998. Bringing context back into epidemiology: Variables and fallacies in multilevel analysis. *American Journal of Public Health,* 88 (2): 216-22.

Duncan, Craig, Kelvyn Jones, and Graham Moon. 1998. Context, composition and heterogeneity: using multilevel models in health research. *Social Science & Medicine,* 46 (1): 97-117.

Enders, Craig K., and Davood Tofighi. 2007. Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods,* 12 (2): 121-38.

Fairbrother, Malcolm. 2011. Explaining social change: the application of multilevel models to repeated cross-sectional survey data. *European Consortium for Political Research General Conference*. Reykjavik. Available at http://www.ecprnet.eu/MyECPR/proposals/reykjavik/uploads/papers/3549.pdf [Accessed 1st March 2012].

Ferron, John M, Kristin Y Hogarty, Robert F Dedrick, Melinda R Hess, Jonhn D Niles, and Jeffrey D Kromrey. 2008. Reporting results from multilevel analysis. Chap. 11 In *Multilevel modeling of educational data*, edited by Ann A O'Connell and Betsy McCoach. 391-426. Charlotte NC: Information Age.

Fielding, Antony. 2004. The role of the Hausman test and whether higher level effects should be treated as random or fixed. *Multilevel modelling newsletter,* 16 (2): 3-9.

Goldstein, Harvey. 2010. *Multilevel statistical models.* 4th edition. Chichester: Wiley.

Green, Donald P., Soo Y. Kim, and David H. Yoon. 2001. Dirty pool. *International Organization,* 55 (2): 441-68.

Greene, William H. 2011a. Fixed Effects Vector Decomposition: A magical solution to the problem of time-invariant variables in fixed effects models? *Political Analysis,* 19 (2): 135-46.

———. 2011b. Fixed-Effects Vector Decomposition: Properties, reliability, and instruments - Reply. *Political Analysis,* 19 (2): 170-72.

———. 2012. *Econometric Analysis.* 7th ed. Harlow: Pearson.

Grilli, Leonardo, and Carla Rampichini. 2011. The role of sample cluster means in multilevel models: A view on endogeneity and measurement error issues. *Methodology-European Journal of Research Methods for the Behavioral and Social Sciences,* 7 (4): 121-33.

Hanchane, Saïd, and Tarek Mostafa. 2011. Solving endogeneity problems in multilevel estimation: an example using education production functions. *Journal of Applied Statistics*.

Hausman, Jerry A. 1978. Specification tests in econometrics. *Econometrica,* 46 (6): 1251-71.

Hausman, Jerry A., and William E. Taylor. 1981. Panel data and unobservable individual effects. *Econometrica,* 49 (6): 1377-98.

Heckman, James J, and Edward Vytlacil. 1998. Instrumental variables methods for the correlated random coefficient model - Estimating the average rate of return to schooling when the return is correlated with schooling. *Journal of Human Resources,* 33 (4): 974-87.

Hsiao, Cheng. 2003. *Analysis of panel data.* Cambridge: CUP.

Jones, Kelvyn. 1991. Specifying and estimating multi-level models for geographical research. *Transactions of the Institute of British Geographers,* NS 16 (2): 148-59.

———. 2005. Random reflections on modelling, geography and voting. Chap. 28 In *Research methods in the social sciences*, edited by Bridget Somekh and Cathy Lewin. 252-55. London: Sage.

———. 2010. The practice of quantitative methods. In *Research Methods in the Social Sciences*, edited by Bridget Somekh and Cathy Lewin. 2nd ed. London: Sage.

Jones, Kelvyn, and Nina Bullen. 1994. Contextual models of urban house prices - a comparison of fixed-coefficient and random-coefficient models developed by expansion. *Economic Geography,* 70 (3): 252-72.

Jones, Kelvyn, and Craig Duncan. 1995. Individuals and their ecologies: analysing the geography of chronic illness within a multilevel modelling framework. *Health and Place,* 1 (1): 27-40.

Jones, Kelvyn, Ron J. Johnston, and Charles J. Pattie. 1992. People, places and regions - exploring the use of multilevel modeling in the analysis of electoral data. *British Journal of Political Science,* 22: 343-80.

Kaldor, Nicholas. 1961. Capital accumulation and economic growth. In *The Theory of capital : proceedings of a conference held by the International Economic Association*, edited by Friedrich A Lutz and Douglas Hague. 177-222. London: Macmillan.

Kennedy, Peter. 2008. *A guide to econometrics.* 6th ed. Malden MA: Blackwell.

King, Gary. 2001. Proper nouns and methodological propriety: Pooling dyads in international relations data. *International Organization,* 55 (2): 497-507.

Kravdal, Øystein. 2011. The fixed-effects model admittedly no quick fix, but still a step in the right direction and better than the suggested alternative. *Journal of Epidemiology and Community Health,* 65 (4): 291-92.

Kreft, Ita, and Jan De Leeuw. 1998. *Introducing multilevel modeling.* London: Sage.

Krishnakumar, Jaya. 2006. Time invriant variables and panel data models: a generalised Frisch-Waugh theorem and its implications. Chap. 5 In *Panel data econometrics: theoretical contributions and empirical applications*, edited by Badi H. Baltagi. 119-32. Amsterdam: Elsevier.

Leckie, George, and Chris Charlton. 2013. runmlwin: A program to run the MLwiN multilevel modelling software from within Stata. *Journal of Statistical Software,* 52 (11).

Leyland, Alastair H. 2010. No quick fix: understanding the difference between fixed and random effect models. *Journal of Epidemiology and Community Health,* 64 (12): 1027-28.

Li, Xiaomei. 2011. Approaches to modelling heterogeneity in longitudinal studies. Victoria University. Available at http://researcharchive.vuw.ac.nz/bitstream/handle/10063/1695/thesis.pdf?sequence=1 [Accessed 26th April 2012].

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.

Moulton, Brent R. 1986. Random Group Effects and the Precision of Regression Estimates. *Journal of Econometrics,* 32 (3): 385-97.
Mundlak, Yair. 1978a. Pooling of time-series and cross-section data. *Econometrica,* 46 (1): 69-85.
———. 1978b. Models with variable coefficients: integration and extension. *Annales de l'inséé,* No 30/31: The Econometrics of Panel Data: 483-509.
Neuhaus, John M., and Jack D. Kalbfleisch. 1998. Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics,* 54 (2): 638-45.
O'Connell, Ann A, and D Betsy McCoach. 2008. *Multilevel modelling of educational data.* Charlotte NC: Information Age.
Oneal, John R., and Bruce Russett. 2001. Clear and clean: The fixed effects of the liberal peace. *International Organization,* 55 (2): 469-85.
Palta, Mari, and Chris Seplaki. 2003. Causes, problems and benefits of different between and within effects in the analysis of clustered data. *Health Services and Outcomes Research Methodology,* 3: 177-93.
Pawson, Ray. 2006. *Evidence-based policy: a realist perspective.* London: Sage.
Plümper, Thomas, and Vera E. Troeger. 2007. Efficient estimation of time-invariant and rarely changing variables in finite sample panel analyses with unit fixed effects. *Political Analysis,* 15 (2): 124-39.
———. 2011. Fixed-Effects Vector Decomposition: Properties, reliability, and instruments. *Political Analysis,* 19 (2): 147-64.
Rasbash, Jon, Chris Charlton, William J Browne, M Healy, and B Cameron. 2013. MLwiN version 2.28. Centre for Multilevel Modelling, University of Bristol.
Rasbash, Jon, Fiona Steele, William J Browne, and Harvey Goldstein. 2009. *A user's guide to MLwiN, version 2.10.* Centre for Multilevel Modelling, University of Bristol.
Raudenbush, Stephen W. 1989. Centering predictors in multilevel analysis: choices and consequences. *Multilevel modelling newsletter,* 2 (1): 10-12.
———. 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. 1986. A hierarchical model for studying school effects. *Sociology of Education,* 59 (1): 1-17.
———. 2002. *Hierarchical linear models: applications and data analysis methods.* 2nd ed. London: Sage.
Rubin, Donald B. 1980. Using empirical Bayes techniques in the law-school validity studies. *Journal of the American Statistical Association,* 75 (372): 801-16.
Schurer, Stefanie, and Jongsay Yong. 2012. Personality, well-being and the marginal utility of income: what can we learn from random coefficient models? *Health, Economics and Data Group, Working Paper*: University of York. Available at http://www.york.ac.uk/res/herc/documents/wp/12_01.pdf [Accessed 16th March 2012].
Shin, Yongyun, and Stephen W. Raudenbush. 2010. A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics,* 35 (1): 26-53.
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.
Skrondal, Anders, and Sophia Rabe-Hesketh. 2004. *Generalized latent variable modeling: multilevel, longitudinal and structural equation models.* Boca Raton: Chapman and Hall.
Snijders, Tom A B, and Johannes Berkhof. 2007. Diagnostic checks for multilevel models. Chap. 3 In *Handbook of Multilevel Analysis*, edited by Jan de Leeuw and Erik Meijer. 139-73. New York: Springer.
Snijders, Tom A B, and Roel J Bosker. 2012. *Multilevel analysis: an introduction to basic and advanced multilevel modelling.* 2nd ed. London: Sage.

Solow, Robert M. 1988. *Growth Theory: An Exposition.* New York: Oxford University Press.
Spanos, Aris. 2006. Revisiting the omitted variables argument: substantive vs statistical adequacy. *Journal of Economic Methodology,* 13 (2): 179-218.
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.
Stegmueller, Daniel. 2013. How many countries do you need for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science,* 57 (3): 748-61.
Subramanian, S. V., Kelvyn Jones, Afamia Kaddour, and Nancy Krieger. 2009a. Revisiting Robinson: The perils of individualistic and ecologic fallacy. *International Journal of Epidemiology,* 38 (2): 342-60.
———. 2009b. Response: The value of a historically informed multilevel analysis of Robinson's data. *International Journal of Epidemiology,* 38 (2): 370-73.
Verbeke, Geert, and Geert Molenberghs. 2000. *Linear mixed models for longitudinal data.* New York: Springer.
———. 2005. *Models for discrete longitudinal data.* New York: Springer.
Wooldridge, Jeffrey M. 2002. *Econometric analysis of cross section and panel data.* Cambridge MA: MIT Press.
———. 2009. Correlated random effects models with unbalanced panels. Michigan State University. Available at http://www.bancaditalia.it/studiricerche/seminari/2011/Wooldridge/paperwooldridge.pdf [Accessed 16th April 2012].
Zorn, Christopher. 2001. Estimating between- and within-cluster covariate effects, with an application to models of international disputes. *International Interactions,* 27 (4): 433-45.

14 Tables

Table 1: Different RE model formulations considered in this paper.

Model Name	Fixed part of model
1. Standard RE	$y_{ij} = \beta_0 + \beta_1 x_{1ij}$
2. Mundlak	$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_3 \bar{x}_j$
3. Within-Between[32]	$y_{ij} = \beta_0 + \beta_1 (x_{ij} - \bar{x}_j) + \beta_4 \bar{x}_j$
4. Within	$y_{ij} = \beta_0 + \beta_1 (x_{ij} - \bar{x}_j)$

[32] The within-between and the within RE model involve group mean centring of the covariate. This is different from centring on the grand mean, which has a different purpose: to keep the value of the intercept (β_0) within the range of the data and to aid convergence of the model. Indeed, x_{1ij} and $\bar{x}_j$ can be grand mean centred if required (the group mean centred variables will already be centred on their grand mean by definition).

Table 2: RMSE, bias and optimism from the simulation results over 5 permutations (times 1000 estimations); Units (30), time periods (20) and the Contextual effect size (1) are kept constant. Correlation between Z3 and uj varies, with values -0.2, 0, 0.2, 0.4 and 0.6. The data are balanced.

	FE	**RE**	**REWB**	**FEVD**
Bias (perfect =1)				
β_3 (within effect of x_{3ij})	0.998	0.978	0.998	0.998
β_6 (effect of t-invariant z_{3j})		1.262	1.261	1.262
β_{3B} (between effect of x_{3ij})			0.969	
RMSE (perfect = 0)				
β_3	0.127	0.130	0.127	0.127
β_6		1.461	1.455	1.463
β_{3B}			0.778	
Optimism (perfect =1)				
β_3	1.007	1.010	1.006	1.007
β_6		1.003	0.975	1.029
β_{3B}			1.004	

Table 3: RMSE, bias and optimism from the simulation results over 5 permutations (times 1000 estimations); Units (30), time periods (20) and the Contextual effect size (1) are kept constant. Correlation between Z3 and uj varies, with values -0.2, 0, 0.2, 0.4 and 0.6. The data are unbalanced.

	FE	**RE**	**REWB**	**FEVD**
Bias (perfect =1)				
β_3 (within effect of x_{3ij})	1.000	0.966	1.000	1.000
β_6 (effect of t-invariant z_{3j})		1.267	1.267	1.267
β_{3B} (between effect of x_{3ij})			0.969	
RMSE (perfect = 0)				
β_3	0.165	0.170	0.165	0.165
β_6		1.484	1.474	1.518
β_{3B}			0.793	
Optimism (perfect =1)				
β_3	0.978	0.987	0.977	0.780
β_6		1.030	1.003	1.333
β_{3B}			1.010	

Table 4: Variables in the trade liberalism analysis, including the amount and proportion of variance that occurs at level 2.

Variable	Explanation	Data Type	Level 2 Variance	VPC
Tariff	Unweighted statutory tariff rate	Continuous	117.220	0.582
Polity	Summary measure of regime type - values between -10 (autocratic) and 10 (democratic) [lagged 1 year]	Ordinal (but treated as continuous)	37.575	0.717
GDPpc	Per capita real GDP [lagged 1 year]	Continuous	1.55e7	0.940
LnPop	Natural Log of population [lagged 1 year]	Continuous	2.281	0.993
Date[33]	Year of tariff measurement	Continuous	1.986	0.071

Table 5: The estimates for trade liberalism analysis

	1. null		2. FE		3. Standard RE (with heterogeneity bias)		4. Within-between RE[34]	
	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.
Fixed Part								
Constant	19.672	1.162	21.954	0.283	21.868	0.960	20.892	0.990
Polity			-0.227	0.086	-0.210	0.076	-0.227	0.086
Lnpop –gm			37.788	6.257	3.322	0.618	37.788	6.240
GDPpc –gm			0.001	0.000	-0.001	0.000	0.001	0.000
Date –gm			-1.813	0.162	-0.996	0.066	-1.813	0.161
Polity mean –gm							-0.055	0.161
Lnpop mean –gm							3.202	0.638
GDPpc mean							-0.001	0.000
Random Part								
Level 2: country								
σ^2_{u0}	117.220	19.093			74.016	12.134	77.838	12.581
Level 1: date								
σ^2_{e0}	84.342	4.590			56.300	3.064	53.581	2.917
-2*loglikelihood:	5854.063				5532.410		5499.771	

[33] With balanced data, the variable Date would have zero level 2 variance. However because of the imbalance of the dataset, there is a small amount of between-variation.

[34] Note that the within estimates were calculated using the variables of the form $(x_{ij} - \bar{x}_j)$ in model 4. Note also that the 'between' means were calculated using the full data, prior to listwise deletion. The within components were calculated using the unit means of the cut down data, to preserve orthogonality.

Table 6: extensions to the RE model for the trade liberalism analysis

	5. with SAsia dummy		6. RCM polity with L1 linear variance		7. With Bangladesh dummy		8. With cross level interaction	
	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.
Fixed Part								
cons	19.004	0.779	19.144	0.784	19.016	0.777	19.096	0.780
Polity W	-0.227	0.086	-0.143	0.187	-0.015	0.132	-0.135	0.138
Lnpop W	37.788	6.247	45.380	6.303	42.916	6.081	40.367	6.155
GDPpc W	0.001	0.000	0.001	0.000	0.001	0.000	0.001	0.000
Date W	-1.813	0.161	-2.000	0.159	-1.942	0.155	-1.901	0.155
Polity mean –gm	-0.203	0.123	-0.227	0.123	-0.246	0.122	-0.203	0.123
Lnpop mean –gm	1.654	0.519	1.687	0.508	1.643	0.498	1.741	0.505
GDPpc mean	-0.001	0.000	-0.001	0.000	-0.001	0.000	-0.001	0.000
[35]SAsia dummy	36.107	4.267	30.566	4.071	33.543	4.038	33.985	4.103
Bangladesh.polity W					-3.605	0.657	-3.614	0.634
Polity W.polity mean							-0.078	0.036
Random Part								
Level 2: country								
σ^2_{u0}	40.281	7.116	41.676	6.984	40.941	6.894	40.836	6.877
σ_{u0u1}			-3.429	1.233	-2.778	0.805	-2.389	0.751
σ^2_{u1} (Polity W)			1.102	0.300	0.378	0.123	0.306	0.106
Level 1: date								
σ^2_{e0}	53.693	2.919	38.820	2.196	39.775	2.237	39.646	2.229
σ_{e0e1}			1.220	0.158	1.223	0.170	1.193	0.181
σ^2_{e1} (PolityW)								
-2*loglikelihood:	5445.315		5284.145		5265.084		5260.55	

[35] SAsia Dummy includes three countries: Bangladesh, India and Pakistan. They could be fitted as a single term (rather than as three separate dummies) without any significant increase in the model deviance.

15 Figures

Figure 1: Plot of level 2 shrunken (intercept) residuals from model 4 of the trade liberalism analysis, with 95% confidence intervals

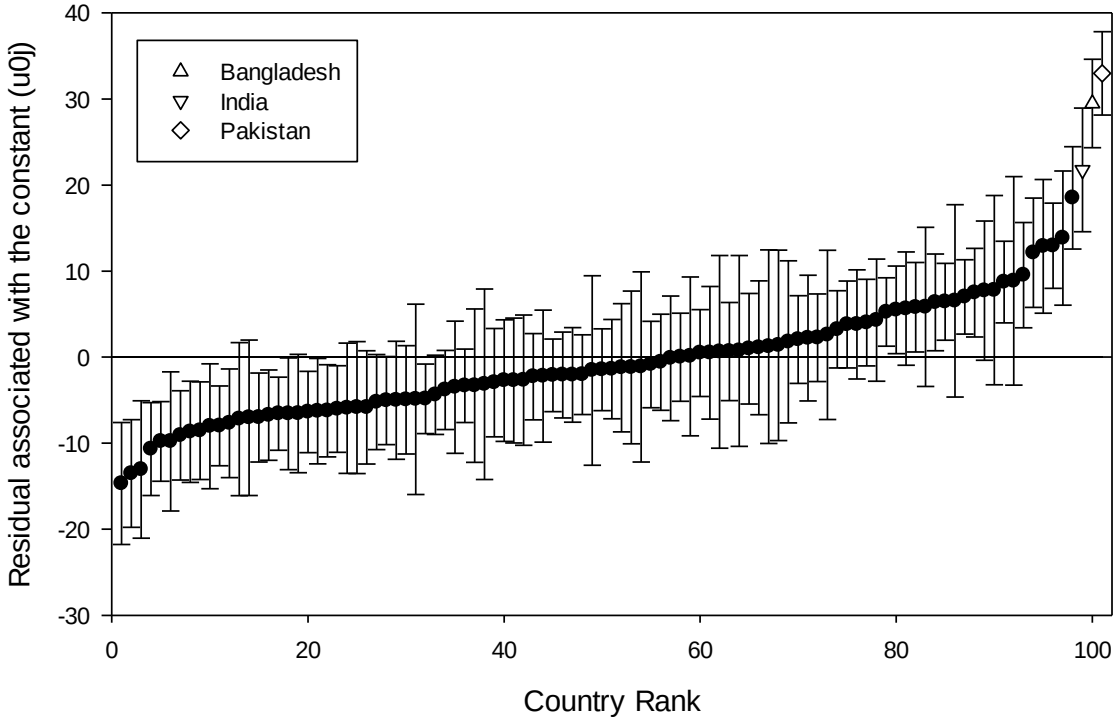

Figure 2: Plot of level 2 random slope shrunken residuals associated with the within polity coefficient from model 6, with 95% confidence intervals.

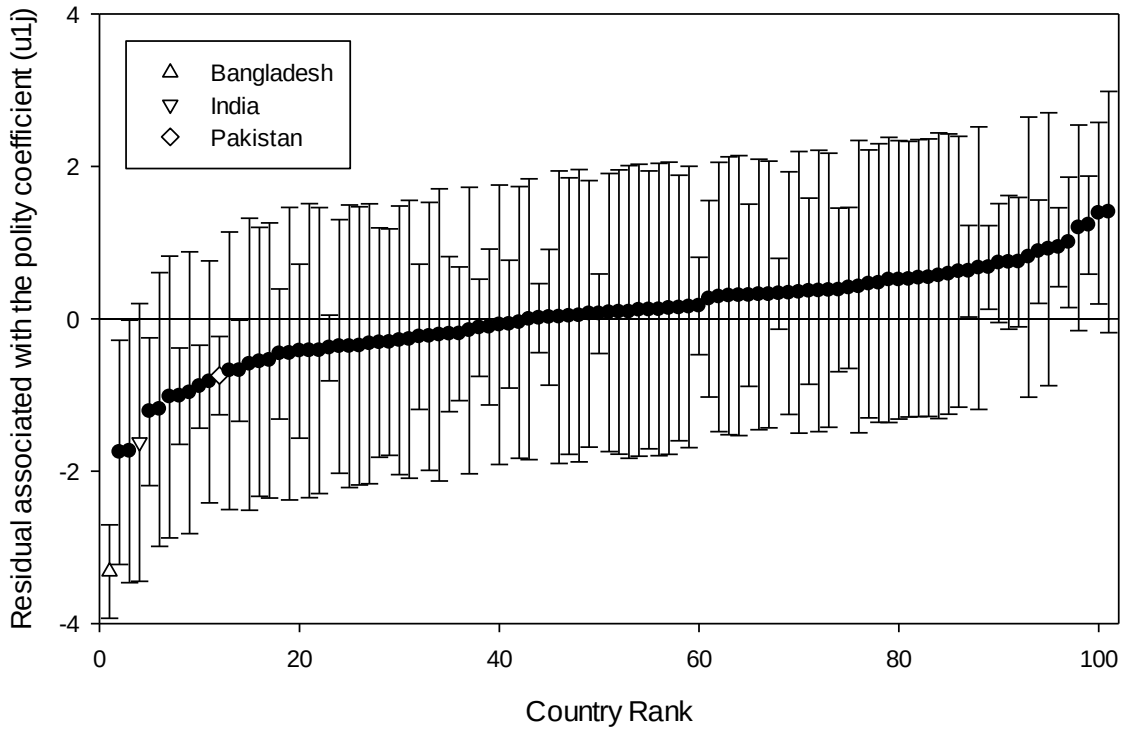

Figure 3: Predictions of the within-effects of polity on each country's tariff rate, from model 7 (with other variables kept constant).

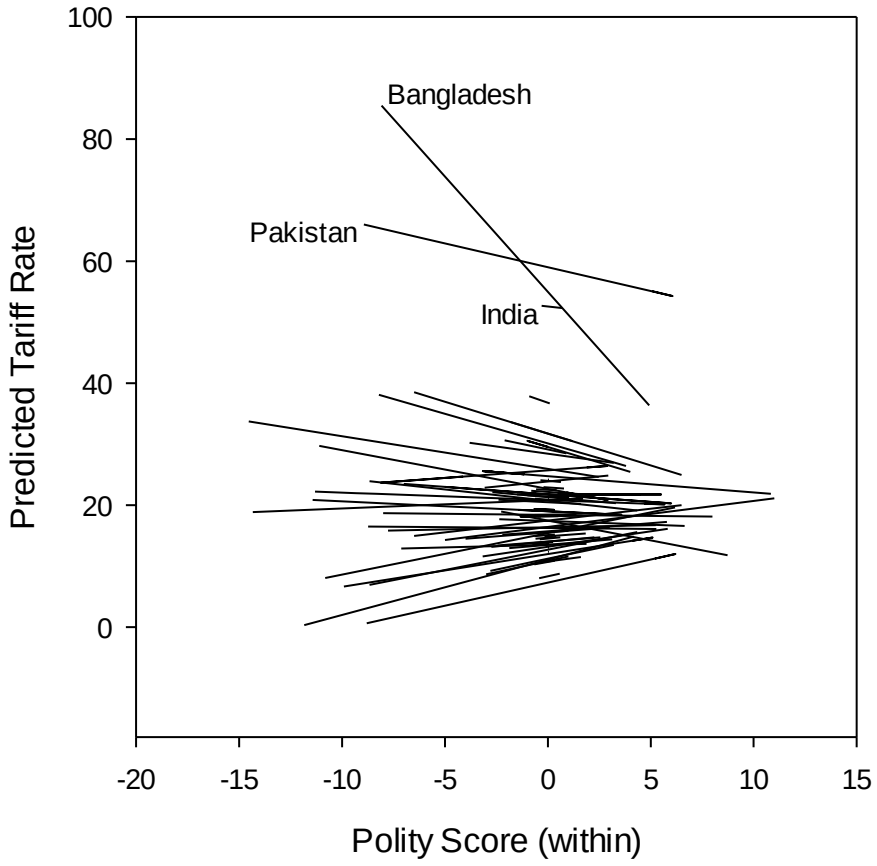

Figure 4: Variance functions at level 1 and level 2 for the within polity effect, from model 7.

With 95% confidence intervals.

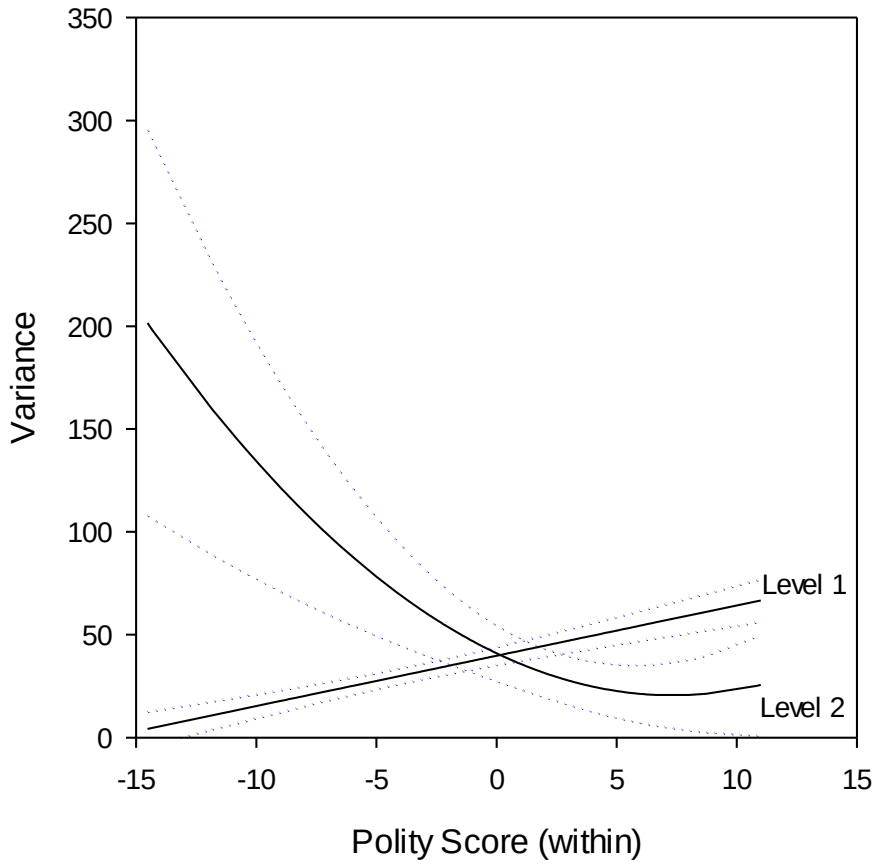

Figure 5: Cross level interaction between the within and between effects of polity, with lines for countries with a mean polity score of +6 and -6 over the period of measurement. With 95% confidence intervals.