Quality & Quantity (2019) 53:1051–1074
https://doi.org/10.1007/s11135-018-0802-x

Fixed and random effects models: making an informed choice

Andrew Bell[1] · **Malcolm Fairbrother**[2] · **Kelvyn Jones**[3]

Published online: 7 August 2018

Abstract
This paper assesses the options available to researchers analysing multilevel (including longitudinal) data, with the aim of supporting good methodological decision-making. Given the confusion in the literature about the key properties of fixed and random effects (FE and RE) models, we present these models' capabilities and limitations. We also discuss the within-between RE model, sometimes misleadingly labelled a 'hybrid' model, showing that it is the most general of the three, with all the strengths of the other two. As such, and because it allows for important extensions—notably random slopes—we argue it should be used (as a starting point at least) in all multilevel analyses. We develop the argument through simulations, evaluating how these models cope with some likely mis-specifications. These simulations reveal that (1) failing to include random slopes can generate anti-conservative standard errors, and (2) assuming random intercepts are Normally distributed, when they are not, introduces only modest biases. These results strengthen the case for the use of, and need for, these models.

Keywords Multilevel models · Fixed effects · Random effects · Mundlak · Hybrid models · Within and between effects

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11135-018-0802-x) contains supplementary material, which is available to authorized users.

✉ Andrew Bell
andrew.j.d.bell@sheffield.ac.uk

Malcolm Fairbrother
Malcolm.fairbrother@umu.se

Kelvyn Jones
kelvyn.jones@bristol.ac.uk

[1] Sheffield Methods Institute, University of Sheffield, ICOSS, 219 Portobello, Sheffield S1 4DP, UK

[2] Sociology Department, Umeå University, Hus Y, Beteendevarhuset, Mediagränd 14, Beteendevetarhuset, Umeå universitet, 901 87 Umeå, Sweden

[3] School of Geographical Sciences and Centre for Multilevel Modelling, University of Bristol, University Road, Bristol BS8 1SS, UK

1 Introduction

Analyses of data with multiple levels, including longitudinal data, can employ a variety of different methods. However, in our view there is significant confusion regarding these methods. This paper therefore presents and clarifies the differences between two key approaches: fixed effects (FE) and random effects (RE) models. We argue that in most research scenarios, a well-specified RE model provides everything that FE provides and more, making it the superior method for most practitioners (see also Shor et al. 2007; Western 1998). However, this view is at odds with the common suggestion that FE is often preferable (e.g. Vaisey and Miles 2017), if not the "gold standard" (e.g. Schurer and Yong 2012). We thus address widespread misunderstandings about FE and RE models, such as those from the literature's use of confusing terminology (including the phrase 'random effects' itself—see for example Gelman 2005) and/or different disciplines' contradictory approaches to the same important methodological questions.

In addition to this synthesis of the inter-disciplinary methodological literature on FE and RE models (information that, whilst often misunderstood, is not new), we present an original simulation study showing how various forms of these models respond in the presence of some plausible model mis-specifications. The simulations show that estimated standard errors are anti-conservative when random-slope variation exists but a model does not allow for it. They also show the robustness of estimation results to mis-specification of random effects as Normally distributed, when they are not; substantial biases are confined to variance and random effect estimates in models with a non-continuous response variable.

The paper begins by outlining what both FE and RE aim to account for: clustering or dependence in a dataset, and differing relationships within and between clusters. We then present our favoured model: a RE model that allows for distinct within and between effects,[1] which we abbreviate "REWB", with heterogeneity modelled at both the cluster (level 2) and observation (level 1) level. Focussing first on the fixed part of the model, we show how the more commonly used FE, RE and pooled OLS models can be understood as constrained and more limited versions of this model; indeed, REWB is our favoured model because of its encompassing nature. Section 3 of this paper focuses on the different treatment of level-2 entities in FE and RE models, and some of the advantages of the RE approach. In Sect. 4, we consider some important extensions to the REWB model that cannot be as effectively implemented under a FE or Ordinary Least Squares framework: 'random slopes' allowing the associations between variables to vary across higher-level entities, further spatial and temporal levels of analysis, and explicit modelling of complex level 1 heteroscedasticity. We show that implementing these extensions can often be of paramount importance and can make results more nuanced, accurate, and informative. Section 5 then considers models with a non-continuous response variable, and some of the distinct challenges that such data present, before considering the assumptions made by the RE model and the extent to which it matters when those assumptions are violated. The article concludes with some practical advice for researchers deciding what model they should use and how.

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

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

Social science datasets often have complex structures, and these structures can be highly relevant to the research question at hand, and not merely a convenience in the research design that has become a nuisance in the analysis. Often, observations (at level 1) are clustered into groups of some kind (at level 2). Such two-level data structures are the main focus of this paper, though data are sometimes grouped at further levels, yielding three (or more) levels. Some of the most common multilevel structures are outlined in Table 1. In broad terms, these can be categorised into two types: cross-sectional data, where individuals are nested within a geographical or social context (e.g. individuals at level 1, within schools or countries at level 2), and longitudinal data, where individuals or social units are measured on a number of occasions. In the latter context, this means occasions (at level 1) are nested within the individual or entity (now at level 2). In all cases these structures represent real and interesting societal configurations; they are not simply a technicality or consequence of survey methodology, as the population may itself be structured by social processes, distinctions, and inequalities.

Structures are important in part because variables can be related at more than one level in a hierarchy, and the relationships at different levels are not necessarily equivalent. Cross-sectionally, for example, some social attitude (Y) may be related to an individual's income X (at level 1) very differently than to the average income in their neighbourhood, country, or region (level 2). A classic example of this comes from American politics. American states with higher incomes therefore tend to elect more Democratic than Republican politicians, but within states richer voters tend to support Republican rather than Democratic candidates (Gelman 2008).

Longitudinally, people might be affected by earning what is, for them, an unusually high annual income (level 1) in a different way than they are affected by being high-earners generally across all years (level 2). The same can hold for whole societies: Europeans for example demand more income redistribution from their governments in times of greater inequality—relative to the average for their country—even though people in consistently more unequal countries do not generally demand more redistribution (Schmidt-Catran 2016). Thus, we can have "within" effects that occur at level 1, and "between" or "contextual" effects that occur at level 2 (Howard 2015), and these three different effects should not be assumed to be the same.

Sometimes it is the case that within effects are of the greatest interest, especially when policy interventions are evaluated. With panel data, for example, within effects can capture the effect of an independent variable changing over time. Many studies have argued for focusing on the longitudinal relationships because unobserved, time-invariant differences between the level 2 entities are then controlled for (Allison 1994; Halaby 2004, see Sect. 2.3). Christmann (2018) for example shows that people are more satisfied with the functioning of democracy in their country during times of good economic performance—a within-country effect that shows the value of improving economic performance.

Yet between effects in longitudinal studies are often equally illuminating, despite being by definition non-changing—as evidenced by the many published studies that rely exclusively on cross-sectional data. Similarly, in cross-sectional studies, the effects of wider social contexts on individuals can also be extremely relevant. Social science is concerned with understanding the world as it exists, not just dynamic changes within it. Thus with a panel dataset for example, it will often be worth modelling associations at the higher level,

Table 1 Some hierarchical structures of data common in social science

Broad category	Data type	Level 1	Level 2	Level 3
Cross-sectional	Clustered survey data (Maimon and Kuhl 2008)	Individuals	Neighbourhoods	–
Cross-sectional	Cross-national survey data (Ruiter and van Tubergen 2009)	Individuals	Countries	–
Cross-sectional	Surveys with multiple items (Deeming and Jones 2015; Sampson et al. 1997)	Items	Individuals	–
Panel	Country time-series cross-sectional data (Beck and Katz 1995; Western 1998)	Occasions	Countries	–
Panel	Individual panel data (Lauen and Gaddis 2013)	Occasions	Individuals	–
Panel at level 1, cross-sectional at level 2	Panel data on individuals who are clustered (Kloosterman et al. 2010)	Occasions	Individuals	Schools
Cross-sectional at level 1, Panel at level 2	Comparative longitudinal survey data (Fairbrother 2014; Schmidt-Catran and Spies 2016), or repeated cross-sectional data (Duncan et al. 1996)	Individuals	Country-years/region-years	Countries/regions

For more elaboration of hierarchical and non-hierarchical structures, see Rasbash (2008)

in order to understand the ways in which individuals differ—not just the ways in which they change over time (see, for example, Subramanian et al. 2009). We take it as axiomatic that we need both micro and macro associations to understand the whole of 'what is going on'.

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

We now outline some statistical models that aim to represent these processes. Taking a panel data example, where individuals i (level 2) are measured on multiple occasions t (level 1), we can conceive of the following model—the most general of the models that we consider in this paper. This specification is able to model both within- and between-individual effects concurrently, and also explicitly models heterogeneity in the effect of predictor variables at the individual level:

$$y_{it} = \mu + \beta_{1W}(x_{it} - \bar{x}_i) + \beta_{2B}\bar{x}_i + \beta_3 z_i + \upsilon_{i0} + \upsilon_{i1}(x_{it} - \bar{x}_i) + \epsilon_{it0} \quad (1)$$

Here y_{it} is the dependent variable, x_{it} is a time-varying (level 1) independent variable, and z_i is a time-invariant (level 2) independent variable. The variable x_{it} is divided into two with each part having a separate effect: β_{1W} represents the average within effect of x_{it}, whilst β_{2B} represents the average between effect of x_{it}.[2] The β_3 parameter represents the effect of time-invariant variable z_i, and is therefore in itself a between effect (level 2 variables cannot have within effects since there is no variation within higher-level entities.) Further variables could be added as required.

The random part of the model includes two terms at level 2—a random effect (υ_{i0}) attached to the intercept and a random effect (υ_{i1}) attached to the within slope—that between them allow heterogeneity in the within-effect of x_{it} across individuals. Each of these are usually assumed to be Normally distributed (as discussed later in this paper).

We will demonstrate in Sect. 4 that specifying heterogeneity at level 2 (with the υ_{i1} term in Eq. 1) can be important for avoiding biases, in particular in standard errors, and this is a key problem with FE and 'standard' RE models. However, to clarify the initial arguments of the first part of this paper, we consider a simplified version of this model that assumes homogeneous effects across level 2 entities:

$$y_{it} = \beta_0 + \beta_{1W}(x_{it} - \bar{x}_i) + \beta_{2B}\bar{x}_i + \beta_3 z_i + \left(\upsilon_i + \epsilon_{it}\right). \quad (2)$$

Here υ_i are the model's (homogeneous) random effects for individuals i, which are assumed to be Normally distributed. The ϵ_{it} are the model's (homoscedastic) level 1 residuals, which are also assumed to be Normally distributed (we will discuss models for non-Gaussian outcomes, with different distributional assumptions, later).

An alternative parameterisation to Eq. 2 (with the same distributional assumptions) is the 'Mundlak' formulation (Mundlak 1978):

$$y_{it} = \beta_0 + \beta_{1W}x_{it} + \beta_{2C}\bar{x}_i + \beta_4 z_i + (\upsilon_i + \epsilon_{it}). \quad (3)$$

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

Here x_{it} is included in its raw form rather than de-meaned form $x_{it} - \bar{x}_i$. Instead of the between effect β_{2B}, the Mundlak model estimates the "contextual effect" β_{2C}. The key difference between these two, as spelled out both graphically and algebraically by Raudenbush and Bryk (2002:140) is that the raw value of the time–varying predictor (x_{it}) is controlled for in the estimate of the contextual effect in Eq. 3, but not in the estimate of the between effect in Eq. 2. Thus if the research question at hand is "what is the effect of a (level 1) individual moving from one level-2 entity to another", the contextual effect (β_{2C}) is of more interest, since it holds the level 1 individual characteristics constant. In contrast, if we simply want to know "what is the effect of changing the level of $\bar{x}_i$, without keeping the level of x_{it} constant?", the between effect (β_{2B}) will provide an answer to that. With longitudinal data, the contextual effect is fairly meaningless: it doesn't make sense for an observation (level 1) to move from one (level 2) individual to another, because they are by definition belonging to a specific individual. It therefore makes little sense to control for those observations in estimating the level 2 effect. As such, the between effect, and thus the REWB model, is generally more informative. When using cross-sectional data, the contextual effect is of interest (since we can imagine level 1 individuals moving between level 2 entities without altering their own characteristics). It can thus measure the additional effect of the level 2 entity, once the individual-level characteristic has been accounted for. The between effect can also be interpreted, but a significant effect could be produced as a result of the composition of level 1 entities, without a country-level construct driving the effect. Note, however, that these models are equivalent, since $\beta_{1W} + \beta_{2C} = \beta_{2B}$; each model conveys the same information and will fit the data equally well and we can obtain one from the other with some simple arithmetic.[3]

In a rare recent example using cross-sectional international survey data, Fairbrother (2016) studied public attitudes towards environmental protection, allowing for separate but simultaneous tests both among and within countries of the associations between key attitudinal variables. This permitted the identification of political trust as an especially critical correlate of greater support for environmental protection at both the individual and national level—an important discovery in the substantive literature.

Both the Mundlak model and the within-between random effects (REWB) models (Eqs. 2 and 3 respectively) are easy to fit in all major software packages (e.g. R, Stata, SAS, as well as more specialist software like HLM and MLwiN). They are simply random effects models with the mean of x_{it} included as an additional explanatory variable (Howard 2015).

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

Having established our 'encompassing' model in its two alternative forms (Mundlak, and within-between), we now present three models that are currently more often used. Showing how each of these is a constrained version of Eqs. 2 or 3 above, we demonstrate the disadvantages of choosing any of them instead of the more general and informative REWB specification.

[3] One potential advantage of the within-between model over the Mundlak specification is that there will be zero correlation between $\bar{x}_i$ and $(x_{it} - \bar{x}_i)$, which can facilitate model convergence. Furthermore, if there is problematic collinearity between multiple $\bar{x}_i$'s, some or all of these can be omitted without affecting the estimates of β_1.

2.2.1 Random effects without within and between separation

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

$$y_{it} = \beta_0 + \beta_1^{RE} x_{it} + \beta_3^{RE} z_i + (v_i + \epsilon_{it}) \tag{4}$$

This approach effectively assumes that $\beta_{1W} = \beta_{2B}$, or equivalently that $\beta_{2C} = 0$, in Eqs. 2 and 3 (Bell et al. 2018). Where this assumption is valid, this model is a good choice, and has benefits over the more general model. Specifically, the estimate of β_1^{RE} will be more efficient than the estimates of β_1 or β_{2B} in Eq. 2, because it can utilise variation at both the higher and lower level (e.g. Fairbrother 2014; Halaby 2004). However, when $\beta_1 \neq \beta_{2B}$, the model will produce a weighted average of the two,[4] which will have little substantive meaning (Raudenbush and Bryk 2002:138). Fortunately, it is easy to test whether the assumption of equal within and between effects is true, by testing the equality of the coefficients in the REWB model), or the significance of the contextual effect in the Mundlak model (for example via a Wald test). If there is a significant difference (and not just that the between effect is significant different from zero) the terms should not be combined, and the encompassing within-between or Mundlak model should be used. This was done by Hanchane and Mostafa (2012) considering bias with this model for school (level 2) and student (level 1) performance. They found that in less selective school systems (Finland), there was little bias and a model like Eq. 4 was appropriate, whilst in more selective systems (UK and Germany) the more encompassing model of Eq. 3 was necessary to take account of schools' contexts and estimate student effects accurately.

This is, in fact, what is effectively done by the oft-used 'Hausman test' (Hausman 1978). Although often (mis)used as a test of whether FE or RE models "should" be used (see Fielding 2004), it is really a test of whether there is a contextual effect, or whether the between and within effects are different. This equates in the panel case to whether the changing within effect (e.g. for an effect of income: the effect of being unusually well paid, such as after receiving a non-regular bonus or a pay rise) is different from the cross-sectional effect (being well paid on average, over the course of the period of observation). Even when within and between effects are slightly different, it may be that the bias in the estimated effect is a price worth paying for the gains in efficiency, depending on the research question at hand (Clark and Linzer 2015). Either way, it is important to test whether the multilevel model in its commonly applied form of Eq. 4 is an uninterpretable blend of two different processes.

2.2.2 Fixed effects model

Depending on the field, perhaps the most commonly used and recommended method of dealing with differing within and between effects as outlined above is 'fixed effects'

[4] Specifically, the estimate will be weighted as: $\beta_{ML} = \frac{w_W \beta_W + w_B \beta_B}{w_W + w_B}$, where w_W is precision of the within estimate, that is $w_W = \left(1 \Big/ \left(SE_{\beta W}\right)^2\right)$ and w_B is precision of the between estimate, $w_B = \left(1 \Big/ \left(SE_{\beta B}\right)^2\right)$. Given the larger sample size (and therefore higher precision) of the within estimate, the model will often tend towards the within estimate.β_W and β_B are the within and between effects, respectively (estimated as β_1 and β_{2B} in Eq. 2), although this would depend on the extent of the unexplained level 1 and 2 variation in the model.

modelling. This approach is equivalent to that represented in Eqs. 2 and 3, except that u_j are specified as fixed effects: i.e. dummy variables are included for each higher-level entity (less a reference category) and the υ_i are not treated as draws from any kind of distribution. The result is that between effects (associations at the higher level) cannot be estimated, and the model can be reduced to:

$$y_{it} = \beta_1(x_{it} - \bar{x}_i) + \left(\upsilon_i + \epsilon_{it}\right). \tag{5}$$

Or reduced even further to:

$$(y_{it} - \bar{y}_i) = \beta_1(x_{it} - \bar{x}_i) + \left(\epsilon_{it}\right). \tag{6}$$

This is the model that most software packages actually estimate, such that they do not estimate the magnitudes of the fixed effects themselves. Thus, the model provides an estimate of the within effect β_1, which is not biased by between effects that are different from them.[5] This is of course what is achieved by the REWB model and the Mundlak model: the REWB model employs precisely the same mean-centring as FE models. However, unlike the REWB and Mundlak specification, the de-meaned FE specification reveals almost nothing about the level-2 entities in the model. This means that many research questions cannot be answered by FE, and it can only ever present a partial picture of the substantive phenomenon represented by the model. With panel data, for example, FE models can say nothing about relationships with independent variables that do not change over time—only about deviations from the mean over time. FE models therefore "throw away important and useful information about the relation between the explanatory and the explained variables in a panel" (Nerlove 2005, p. 20).

If a researcher has no substantive interest in the between effects, their exclusion is perhaps unimportant, though even in such a case, for reasons discussed below, we think there are still reasons to disfavour the FE approach as the one and only valid approach. To be clear the REWB and Mundlak will give exactly the same results for the within effect (coefficient and standard error) as the FE model (see Bell and Jones 2015 for simulations; Goetgeluk and Vansteelandt 2008 for proof of consistency), but retains the between effect which can be informative and cannot be obtained from a FE model.

2.2.3 Single level OLS regression

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

$$y_{it} = \beta_0 + \beta_1^{OLS} x_{it} + \beta_4^{OLS} z_i + \left(\epsilon_{it}\right) \tag{7}$$

Thus, we assume that all observations in the dataset are conditionally independent. This has two problems. First, as with the standard RE model, the estimate of β_1^{OLS} will be a potentially uninterpretable weighted average[6] of the within and between effects (if they are not equal). Furthermore, if there are differences between level 2 entities (that is, if there are effects of unmeasured higher-level variables), standard errors will be estimated

[5] Note though that, in the longitudinal setting, between effects will only be fully controlled if those effects do not change over time (this is the case with the REWB/Mundlak models as well, unless such heterogeneity is explicitly modelled).

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

as if all observations are independent, and so will be generally underestimated, especially for parameters associated with higher-level variables, including between and contextual effects.[7] Fortunately, the necessity of modelling the nested structure can readily be evaluated, by running the model both with and without the higher-level random effects and testing which is the better fitting model by a likelihood ratio test (Snijders and Bosker 2012:97), AIC, or BIC.

2.3 Omitted variable bias in the within-between RE model

We hope the discussion above has convinced readers of the superiority of the REWB model, except perhaps when the within and between effects are approximately equal, in which case the standard RE model (without separated within and between effects) might be preferable for reasons of efficiency.[8] Even then, the REWB model should be considered first, or as an alternative, since the equality of the within and between coefficients should not be assumed. As for FE, except for simplicity there is nothing that such models offer that a REWB model does not.

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

As with fixed effects models, the REWB specification prevents any bias on level 1 coefficients due to omitted variables at level 2. To put it another way, there can be no correlation between level 1 variables included in the model and the level 2 random effects—such biases are absorbed into the between effect, as confirmed by simulation studies (Bell and Jones 2015; Fairbrother 2014). When using panel data with repeated measures on individuals, unchanging and/or unmeasured characteristics of an individual (such as intelligence, ability, etc.) will be controlled out of the estimate of the within effect. However, unobserved time-varying characteristics can still cause biases at level 1 in either an FE or a REWB/Mundlak model. Similarly, in a REWB/Mundlak models, unmeasured level 2 characteristics can cause bias in the estimates of between effects and effects of other level 2 variables.

This is a problem if we wish to know the direct causal effect of a level 2 variable: that is, what happens to Y when a level 2 variable increases or decreases, such as because of an intervention (Blakely and Woodward 2000). However, this does not mean that those estimated relationships are worthless. Indeed, often we are not looking for the direct, causal effect of a level 2 variable, but see these variables as proxies for a range of unmeasured social processes, which might include those omitted variables themselves. As an example, in a panel data structure when considering the relationship between ethnicity (an unchanging, level 2 variable) and a dependent variable, we would not interpret any association found to be the direct causal effect of any particular genes or skin pigmentation; rather we

[7] One could add a group mean variable to this equation, as in Eqs. 2 or 3. Whilst this would solve the issue of bias of the point estimates, standard errors would still be underestimated.

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

are interested in the effects of the myriad of unmeasured social and cultural factors that are related to ethnicity. If a direct genetic effect is what we are looking for, then our estimates are likely to be 'biased', but we hope most reasonable researchers would not interpret such coefficients in this way. As long as we interpret any coefficient estimates with these unmeasured variables in mind, and are aware that such reasoning is as much conceptual and theoretical as it is empirical, such coefficients can be of great value in helping us to understand patterns in the world through a model-based approach. Note that if we are, in fact, interested in a direct causal effect and are concerned by possible omitted variables, then instrumental variable techniques can sometimes be employed within the RE framework (for example, see Chatelain and Ralf 2018; Steele et al. 2007).

The logic above also applies to estimates of between and contextual effects. These aggregated variables are proxies of group level characteristics that are to some extent unmeasured. As such, it is not a problem in our view that, in the case of panel data, future data is being used to form this variable and predict past values of the dependent variable—these values are being used to get the best possible estimate of the unchanging group-level characteristic. If researchers want these variables to be more accurately measured, they could be precision-weighted, to shrink them back to the mean value for small groups (Grilli and Rampichini 2011; Shin and Raudenbush 2010).

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

This section aims to clarify further the statistical and conceptual differences between RE (including REWB) and FE modelling frameworks. The obvious difference between the two models is in the way that that the level-2 entities are treated: that is υ_i in Eqs. 2 and 5.

In a RE model (whether standard, REWB or Mundlak) level-2 random effects are treated as random draws from a Normal distribution, the variance of which is estimated:

$$\upsilon_i \sim N\left(0, \sigma_\upsilon^2\right). \tag{8}$$

In contrast, a FE model treats level-2 entities as unconnected: υ_i in Eq. 5 are dummy variables for higher-level entity i, each with separately estimated coefficients (less a reference category, or with the intercept suppressed). Because these dummy variables account for all the higher-level variance, no other variables measured at the higher level can be identified.

In both specifications, the level-1 variance is typically assumed to follow a Normal distribution:

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

To us, this is what the 'random' and 'fixed' in RE and FE mean. In contrast, others argue that *the defining feature* of the RE model is an assumption that that model makes. Vaisey and Miles (2017:47) for example state:

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

Such views are also characteristic of mainstream econometrics:

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

No doubt this assumption is important (see Sect. 2.3). But regardless of how well established this definition is, it is misleading. This assumption is not the only difference between RE and FE models, and is far from being either model's defining feature.

The different distributional assumptions affect the extent to which information is considered exchangeable between higher-level entities: are they unrelated, or is the value of one level-2 entity related to the values of the others? In the FE framework, nothing can be known about each level-2 entity from any or all of the others—they are unrelated and each exist completely independently. At the other extreme, a single-level model assumes there are no differences between the higher-level entities, in a sense knowing one is sufficient to know them all. RE models strike a balance between these two extremes, treating higher-level entities as distinct but not completely unlike each other. In practice, the random intercepts in RE models will correlate strongly with the fixed effects in a 'dummy variable' FE models, but RE estimates will be drawn in or 'shrunk' towards their mean—with unreliably estimated and more extreme values shrunk the most.

Why does it matter that the random effects are drawn from a common distribution? We have already stated that FE models estimate coefficients on higher-level dummy variables (the fixed effects), and cannot estimate coefficients on other higher-level variables (between effects). RE models can yield estimates for coefficients on higher-level variables because the random effects are parameterised as a distribution instead of dummy variables. Moreover, RE automatically provides an estimate of the level 2 variance, allowing an overall measure of the extent to which level-2 entities differ in comparison to the level 1 variance. Further, this variance can be used to produce 'shrunken' (or 'Empirical Bayes') higher-level residuals which, unlike FE dummy-variable parameter estimates, take account of the unreliability of those estimates; for an application, see Ard and Fairbrother (2017). The degree of "shrinkage" (or exchangeability across level 2 entities) in a RE model is determined from the data, with more shrinkage if there are few observations and/or the estimated variance of the level-2 entities, σ_v^2, is small (see Jones and Bullen 1994; Spiegelhalter 2004).

If we are interested in whether individuals' responses are related to their *specific* contexts (neighbourhoods, schools, countries, etc.) a fixed effects model can help answer this question if dummy variables for level-2 entities are estimated, but this is done unreliably with small level-2 entities. A RE model can give us more reliable, appropriately conservative estimates of this (Bell et al. 2018), as well as telling us whether that context matters *in general*, based on the significance of the estimated variance of the random effects.[9] It can tell us both differences in higher-level effects (termed 'type A' effects in the education

[9] This could also be done on the basis of a Wald test of the joint significance of FE dummy variables, but this is not possible with non-linear outcomes where dummy coefficients are not estimated.

literature, Raudenbush and Willms 1995) and the effects of variables at the higher level ('type B' effects). FE estimators cannot estimate the latter.

The view of FE and RE being defined by their assumptions has led many to characterise the REWB model as a 'hybrid' between FE and RE, or even a 'hybrid FE' model (e.g. Schempf et al. 2011). We hope the discussion above will convince readers that this model is a RE model. Indeed, Paul Allison, who (we believe) introduced the terminology of the Hybrid model (Allison 2005, 2009) now prefers the terminology of 'within-between RE' (Allison 2014).

The label matters, because FE models (and indeed 'hybrid' models) are often presented as a technical solution, following and responding to a Hausman test taken to mean that a RE model cannot be used.[10] As such, researchers rarely consider what problem FE actually solves, and *why* the RE parameter estimates were wrong. This bias is often described as 'endogeneity', a term that covers a wide and disparate range of different model misspecifications (Bell and Jones 2015:138). In fact, the Hausman test simply investigates whether the between and within effects are different—a possibility that the REWB specification allows for. REWB (a) recognises the possibility of differences between the within and between effects of a predictor, and (b) *explicitly models* those separate within and between effects. The REWB model is a direct, substantive solution to a mis-specified RE model in allowing for the possibility of different relations at each level; it models between effects, which may be causing the problem, and are often themselves substantively interesting. When treated as a FE model, this substance is often lost.

Further, using the REWB model as if it were a FE model leads researchers to use it without taking full advantage of the benefits that RE models can offer. The RE framework allows a wider range of research questions to be investigated: involving time-invariant variables, shrunken random effects, additional hierarchical (e.g. geographical) levels and, as we discuss in the next section, random slopes estimates that allow relationships to vary across individuals, or allow variances at any level to vary with variables. As well as yielding new, substantively interesting results, such actions can alter the average associations found. Describing the REWB, or Hybrid, model as falling under a FE framework therefore undersells and misrepresents its value and capabilities.

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

4.1 Random slopes models

So far, all models have assumed homogeneity in the within effect associated with x_{it}. This is often a problematic assumption. First, such models hide important and interesting heterogeneity. And second, estimates from models that assume homogeneity incorrectly will suffer from biased estimates, as we show below. The RE/REWB model as previously described also suffers from this shortcoming, but can more easily avoid it by explicitly modelling such heterogeneity, with the inclusion of random slopes (Western

[10] Many (e.g. Greene 2012:421) even argue that the Mundlak or REWB model can be used as a form of the Hausman test, which could be itself be used to justify the use of FE, even though the REWB model makes that choice unnecessary.

Table 2 Results from reanalyses of Milner and Kubota (2005) and Reinhart and Rogoff (2010)

Original study/studies	Milner and Kubota (2005)	Reinhart and Rogoff (2010) and Herndon et al. (2014)
Reanalysis	Bell and Jones (2015) (appendix)	Bell et al. (2015)
Dependent variable	Tariff rates	Economic growth (ΔGDP)
Independent variable of interest	Democracy (polity score)	National debt (%GDP)
REWB/FE within estimate (SE)	−0.227 (0.086)**	−0.021 (0.003)***
Random slopes estimate (SE)	−0.143 (0.187) (NS)	−0.021 (0.009)*
Notes	Effect further reduced by the removal of a single outlying country, Bangladesh.	Effect becomes insignificant when time is appropriately controlled.

Standard errors are in parentheses

For full details of the models used, see the reanalysis papers themselves

NS not significant

P values *** < 0.001; ** < 0.01; * < 0.05

1998). These allow the coefficients on lower-level covariates to vary across level-2 entities. Equation 2 then becomes:

$$y_{it} = \beta_0 + \beta_{1W}(x_{it} - \bar{x}_i) + \beta_{2B}\bar{x}_i + \beta_3 z_i + \upsilon_{i0} + \upsilon_{i1}(x_{it} - \bar{x}_i) + \epsilon_{it} \tag{10}$$

Here β_{1W} is a weighted average (Raudenbush and Bloom 2015) of the within effects in each level-2 entity; υ_{i1} measures the extent to which these within effects vary between level-2 entities (such that each level-2 entity i has a within effect estimated as $\beta_{1W} + \upsilon_{i1}$). The two random terms υ_{i1} and υ_{i0} are assumed to be draws from a bivariate Normal distribution, meaning Eq. 8 is extended to:

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

The meaning of individual coefficients can vary depending on how variables are scaled and centred. However, the covariance term indicates the extent of 'fanning in' (with negative $\sigma_{\upsilon 01}$) or 'fanning out' (positive $\sigma_{\upsilon 01}$) from a covariate value of zero (Bullen et al. 1997). In many cases, there is substantive heterogeneity in the size of associations among level-2 entities. Table 2 shows two examples of reanalyses where including random coefficients makes a real difference to the results. Both are analyses of countries, rather than individuals, but the methodological issues are similar. The first is a reanalysis of an influential study in political science (Milner and Kubota 2005) which claims that political democracy leads to economic globalisation (measured by countries' tariff rates). When including random coefficients in the model, not only does the overall within effect disappear, but a single outlying country, Bangladesh, turns out to be driving the relationship (Bell and Jones 2015, Appendix). The second example is the now infamous study in economics by Reinhart and Rogoff (2010), which claimed that higher levels of public debt cause lower national economic growth (a conclusion that remained even after the Herndon et al. (2014) corrections). In this case, although the coefficient does not change with the introduction of random slopes, the standard error triples in size, and the within effect is no longer statistically significant when, in addition, time is appropriately controlled (Bell et al. 2015).

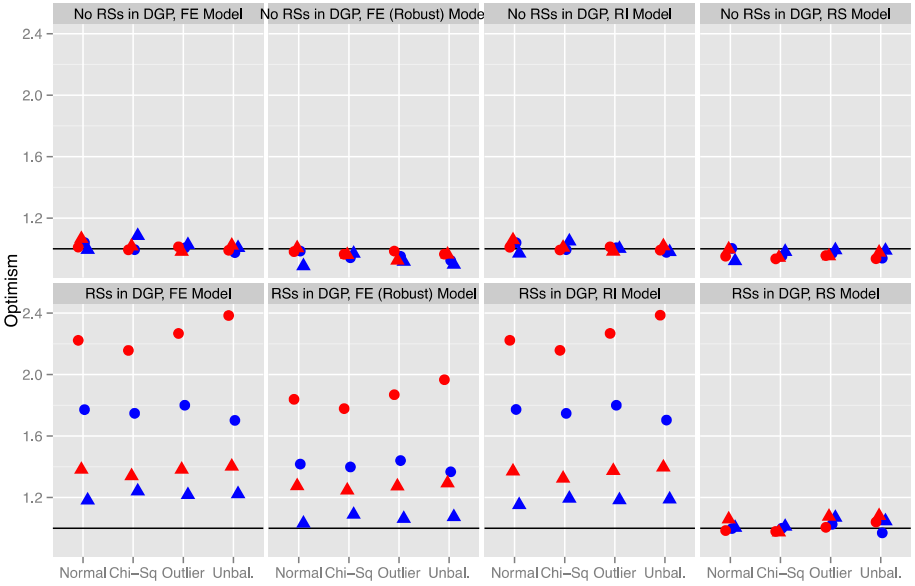

Fig. 1 Optimism of the standard errors in various models. *Note* Triangles are for logistic models, circles for Normal models; blue means 60 groups of ten, red 30 groups of 20. (Color figure online)

In both cases, not only is substantively interesting heterogeneity missed in models assuming homogenous associations, but also within effects are anticonservative (that is, SEs are underestimated). Leaving aside the substantive interest that can be gained from seeing how different contexts can lead to different relationships, failing to consider how associations differ across level-2 entities can produce misleading results if such differences exist. Although such heterogeneity can be modelled in a FE framework with the addition of multiple interaction terms, it rarely is in practice, and that heterogeneity does not benefit from shrinkage as in the RE framework. Thus, a FE model can lead an analyst to miss problematic assumptions of homogeneity that the model is making. A RE model—including the REWB model—allows for the modelling of important complexities, such as heterogeneity across level-2 entities.

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

Figure 1 shows the 'optimism'—the ratio of the true sampling variability to the sampling variability estimated by the standard error (see Shor et al. 2007)—for a single covariate, in a variety of scenarios.[11] In the scenarios presented in the top row, the DGP included only random intercepts, not random slopes; the lower row represents DGPs with both

[11] See the "Appendix" of the present paper for the full explanation and R code to replicate these simulations in ESM.

random intercepts and random slopes. FE models are in the first two columns (with naïve and robust standard errors), random-intercepts models the third column, and random slopes models in the right-hand column.

Figure 1 shows that where random slopes are not included in the analysis model (all but the right-most column), but exist in the data in reality (bottom row), the standard errors are overoptimistic—they are too small relative to the true sampling variability. When there is variation in the slopes across level-2 entities, there is more uncertainty in the beta estimates, but this is not reflected in the standard error estimates unless those random slopes are explicitly specified. In the top row, in contrast, all four columns look the same: here there is no mismatch between the invariant relationships assumed by the analysis models and present in the data. In the presence of heterogeneity, note that while FE models with naive SEs are the most anticonservative, neither FE models with "robust" standard errors nor RE models with only random intercepts are much better.

These results support the strong critique by Barr et al. (2013) that not to include random slopes is anticonservative. On the other hand, Matuschek et al. (2017) counter that analytical models should also be parsimonious, and fitting models with many random effects quickly multiplies the number of parameters to be estimated, particularly since random slopes are generally given covariances as well as variances. Sometimes the data available will not be sufficient to estimate such a model. Still, it will make sense in much applied work to test whether a statistically significant coefficient remains so when allowed to vary randomly. We discuss this further in the conclusions.

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

Datasets often have structures that span more than two levels. A further advantage of the multilevel/random effects framework over fixed effects is its allowing for complex data structures of this kind. Fixed effects models are not problematic when additional higher levels exist (insofar as they can still estimate a within effect), but they are unable to include a third level (if the levels are hierarchically structured), because the dummy variables at the second level will automatically use up all degrees of freedom for any levels further up the hierarchy. Multilevel models allow competing explanations to be considered, specifically at which level in a hierarchy matters the most, with a highly parsimonious specification (estimating a variance parameter at each level).[12]

For example, cross-national surveys are increasingly being fielded multiple times in the same set of countries, yielding survey data that are both comparative and longitudinal. This presents a three-level hierarchical structure, with observations nested within country-years, which are in turn nested in countries (Fairbrother 2014).[13]

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

[13] See Schmidt-Catran and Fairbrother (2015) for a further extension that includes a cross-classified year or wave level.

4.3 Complex level 1 heterogeneity

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

$$y_{it} = \mu + \beta_1(x_{it} - \bar{x}_i) + \beta_2\bar{x}_i + \beta_3 z_i + \upsilon_{i0} + \upsilon_{i1}(x_{it} - \bar{x}_i) + \epsilon_{it0} + \epsilon_{it1}(x_{it} - \bar{x}_i), \quad (12)$$

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

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

Often this is important to do, because what is apparent higher-level variance[14] between level-2 entities, can in fact be complex variance at level 1. It is only by specifying both, as in Eq. 12, that we can be sure how variance, and varying variance, can be attributed between levels (Vallejo et al. 2015).[15]

5 Generalising the RE model: binary and count dependent variables

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

Unlike models using the identity link function, results using the REWB model with other link functions do not produce results that are identical to FE (or the equivalent conditional likelihood model). In other words, the inclusion of the group mean in the model does not reliably partition any higher-level processes from the within effect, meaning both within and between estimates of cluster-specific effects[16] can be biased. This is the case when the relationship between the between component of X ($\bar{x}_i$) and the higher-level residual (υ_i) is non-linear. How big a problem is this? Brumback et al. (2010:1651) found that, in running simulations, "it was difficult to find an example in which the problem is severe"

[14] Note the random slopes described in 4.1 can also be conceived as varying variance. Variance could vary by both level 1 and level 2 variables. The approach used here is standard in the multilevel literature (Goldstein 2010), but other approaches are possible (for example modelling the log of the variance as a function of covariates - e.g. see Hedeker and Mermelstein 2007).

[15] Although difficult to implement in some standard software packages (it cannot be implemented in the mixed package in Stata, or lme4 in R), it can be implemented in MLwiN, which can in turn be accessed from Stata/R using the packages runmlwin/R2MLwiN (Leckie and Charlton 2013; Zhang et al. 2016).

[16] Note: we do not consider the differences between population average and cluster specific estimates in this paper—all models considered in this section of the paper produce the latter. This debate is beyond the scope of the paper (but see Jones and Subramanian 2013; Subramanian and O'Malley 2010 for more on this). Both cluster specific and population average estimates may be needed depending on the research question; this is not a debate that can or should be technically resolved.

(see also Goetgeluk and Vansteelandt 2008). In a later paper, however, Brumback et al. (2013) did identify one such example, but only with properties unlikely to be found in real life data (Allison 2014)—$\bar{x}_i$ and v_i very highly correlated, and few observations per level-2 entity.

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

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

A key assumption of RE models is that the random effects representing the level-2 entities are drawn from a Normal distribution. However, "the Normality of [the random coefficients] is clearly an assumption driven more by mathematical convenience than by empirical reality" (Beck and Katz 2007:90). Indeed, it is often an unrealistic assumption, and it is important to know the extent to which different estimates are biased when that assumption is broken.

The evidence from prior simulations studies is somewhat mixed, and depends on what specifically in the RE model is of interest. For linear models with a continuous response variable, and on the positive side, Beck and Katz (2007) find that both average parameter estimates and random effects are well estimated, both when the random effects are assumed to be Normally distributed but in fact have a Chi square distribution, or there are a number of outliers in the dataset.[17] Others concur that beta estimates are generally unbiased by non-Normal random effects, as are estimates of the random effects variances (Maas and Hox 2004; McCulloch and Neuhaus 2011a). Random effects are only biased to a significant degree in extreme scenarios (McCulloch and Neuhaus 2011b), and even then (for example for random effects with a Chi square(1) distribution), the ranked order of estimated random effects remains highly correlated (Correlation > 0.8) to the rankings of the true random effects (Arpino and Varriale 2010), meaning substantive interpretation is likely to be affected only minimally. This is the case whether or not the DGP includes random slopes. In other words, a badly specified random distribution may result in some biases, but these are usually small enough not to worry the applied researcher. If there is a concern about

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

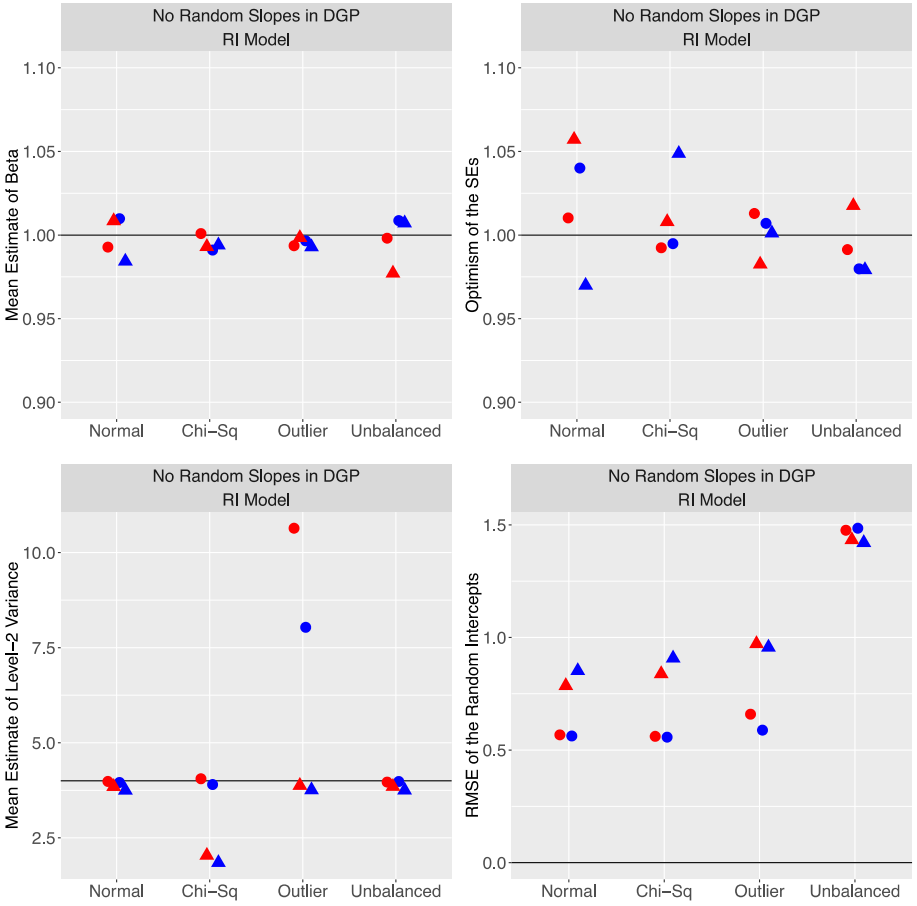

Fig. 2 Biases and RMSE under various (mis-)specifications. *Note* Triangles are for logistic models, circles for Normal models; blue means 60 groups of ten—red 30 groups of 20. Clockwise from the upper-left, the parameters are beta (bias), optimism of the standard errors (bias), random intercepts (RMSE), and level-2 variance (bias). (Color figure online)

bias, it may be wise to check the findings are robust to other specifications, and potentially use models that allow for non-Normal random effects, such as Non-Parametric Maximum Likelihood techniques (Aitkin 1999; Fotouhi 2003).

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

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

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

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

7 Conclusion: what should researchers do?

We hope that this article has presented a clear picture of the key properties, capabilities, and limitations of FE and RE models, including REWB models. We have considered what each of these models are, what they do, what they assume, and how much those assumptions matter in different real-life scenarios.

There are a number of practical points that researchers should take away from this paper. First and perhaps most obviously is that the REWB model is a more general and encompassing option than either FE or conventional RE, which do not distinguish between within and between effects. Even when using non-identity link functions, or when the Normality assumption of the random effects is violated, the small biases that can arise in such models will often be a price worth paying for the added flexibility that the REWB model provides. This is especially the case since FE is unable to provide any estimates at all of the parameters that are most biased by violations of Normality (specifically random effects and variance estimates). The only reason to choose FE is if (1) higher-level variables are of no interest whatsoever, (2) there are no random slopes in the true DGP, or (3) there are so few level-2 entities that random slopes are unlikely to be estimable. Regarding (1) we would argue this is rarely the case in social science, where a full understanding of the world and how it operates is often the end goal. Regarding (2), testing this requires fitting a RE model in any case, so the benefits of reverting to FE are moot. Regarding (3), the REWB model will still be robust for fixed-part parameter estimates (although maximum likelihood estimation may be biased—McNeish 2017; Stegmueller 2013), though it's efficacy relative to FE would be very limited, since higher level parameters would be estimated with a lot of uncertainty.

Second, the question of whether to include random slopes is important and requires careful consideration. On the one hand, in a world of limited computing power and limited data, it is often not feasible to allow the effects of all variables to vary between level-2 entities. On the other hand, we have shown that results can change in substantive ways when slopes are allowed to vary randomly. We would argue that, at the least, where there is a single substantive predictor variable of interest, it would make sense to check that the

conclusions hold when the effect of that variable is allowed to vary across clusters. One option in this regard is to use robust standard errors, not as a correction per se, but as a diagnostic procedure—a 'canary down the mine'—following King and Roberts (2015). Any difference between conventional and robust standard errors suggests there is some kind of misspecification in the model, and that misspecification might well include the failure to model random slopes. The two leftmost panels in the lower row in Fig. 1 show precisely how robust standard errors will differ when a model is mis-specified in omitting relevant random effects.

Third, and in contrast to much of the applied literature, we argue that researchers should not use a Hausman test to decide between fixed and random effects models. Rather, they can use this test, or models equivalent to it, to verify the equivalence of the within and between relationships. A lack of equality should be in itself of interest and worthy of further investigation through the REWB model.

Appendix: The simulations

We generated datasets according to the formula

$$y_{it} = \beta_0 + \beta_1 x_{it} + \upsilon_{i0} + \epsilon_{it},$$

or in other words with random intercepts only, and also according to

$$y_{it} = \beta_0 + \beta_1 x_{it} + \upsilon_{i0} + \upsilon_{i1} x_{it} + \epsilon_{it}.$$

In this latter case, the data-generated process (DGP) included both random intercepts and random slopes, and these random effects were distributed according to

$$\begin{bmatrix} \upsilon_{i0} \\ \upsilon_{i1} \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma^2_{\upsilon 0} & \\ 0 & \sigma^2_{\upsilon 1} \end{bmatrix}\right).$$

That is, the random effects were in all cases uncorrelated. We also generated binary data based on similar models (both random intercept-only and random intercept, random slope models), using a logit link. In all cases, $\sigma^2_{\upsilon 0}$ and $\sigma^2_{\upsilon 1}$ were set to 4, and (for the Normally distributed data) the variance of ϵ_{it} to 1. The overall intercept β_0 and the overall slope β_1 were also set to 1. The x_{it} data were drawn from a Normal distribution with a mean of 0 and a variance of 0.25^2.

We fitted models to simulated data sets with either 60 groups of 10 or 30 groups of 20, yielding a total N of 600 either way.[18] The 30×20 condition reflected that time-series cross-sectional datasets often possess roughly those N's at each level, and that many cross-national survey datasets include about 30 countries. The 60x10 condition allowed for a useful contrast testing the implications of varying the N at either level. We did not conduct simulations with groups larger than 20 because of the high time costs of doing so, and

[18] The N's at each level are not typical of published studies using multilevel models. But most studies use large N's that would have made the simulation studies much more time-consuming to run, with no benefit in terms of insights.

because previous simulation studies have not revealed anything particularly notable about studies conducted with large rather than small groups (Bryan and Jenkins 2016; Schmidt-Catran and Fairbrother 2015).

In some cases, instead of drawing the υ_{i0}'s from a Normal distribution, we drew them from a Chi squared distribution, or from a Normal distribution but with a single large outlier. Where they were drawn from a Chi squared distribution, the distribution's degrees of freedom was set at 2, and we also subtracted 2 from each randomly drawn value, yielding a final population mean of 0 and variance of 4—the same as in scenarios where the υ_{i0}'s were drawn from a Normal distribution. For the scenarios with the outlier, we tripled the value of the element of υ_{i0} with the largest absolute value.

As a fourth possibility, we made the simulated dataset unbalanced, by resampling with replacement a dataset of the same total size from the values of the original, with equal probability of selection. This yielded groups of randomly varying sizes.

In sum, under each of these four conditions (Normal, Chi squared, outlier, unbalanced), we simulated datasets using only random intercepts or both random intercepts and random slopes, with y either Normal or binary, and with one combination of N's or the other—yielding 32 distinct DGPs ($4\times2\times2\times2$). We conducted 1000 simulations with each DGP.

We then fitted three different models to each simulated dataset: a fixed effects model (with naïve and clustered standard errors), a random intercepts-only model, and a random intercepts-random slopes model.

We conducted the simulations in R. For fitting multilevel models we used the package lme4 (Bates et al. 2015). For deriving clustered standard errors from the fixed effects models, we used the plm package (Croissant and Millo 2008). We caught false or questionable convergences and simply removed them, simulating a new dataset instead (this should not bias the results, although it should be noted as an advantage of FE is that it is unlikely to show convergence problems due to being estimated by OLS). We tried multiple runs of simulations, and found stable results beyond about 200 simulations per DGP.

References

Aitkin, M.: A general maximum likelihood analysis of variance components in generalized linear models. Biometrics **55**(1), 117–128 (1999)

Allison, P.D.: Using panel data to estimate the effects of events. Sociol. Methods Res. **23**(2), 174–199 (1994)

Allison, P.D.: Fixed Effects Regression Methods for Longitudinal Data using SAS. SAS Press, Cary, NC (2005)

Allison, P.D.: Fixed Effects Regression Models. Sage, London (2009)

Allison, P.D.: Problems with the hybrid method. Stat. Horiz. http://www.statisticalhorizons.com/problems-with-the-hybrid-method (2014). Accessed 16 July 2015

Ard, K., Fairbrother, M.: Pollution prophylaxis? social capital and environmental inequality*. Soc. Sci. Q. **98**(2), 584–607 (2017)

Arpino, B., Varriale, R.: Assessing the quality of institutions' rankings obtained through multilevel linear regression models. J. Appl. Econ. Sci. **5**(1), 7–22 (2010)

Barr, D.J., Levy, R., Scheepers, C., Tily, H.J.: Random effects structure for confirmatory hypothesis testing: keep it maximal. J. Mem. Lang. **68**(3), 255–278 (2013)

Bates, D., Mächler, M., Bolker, B., Walker, S.: Fitting linear mixed-effects models using lme4. J. Stat. Softw. **67**(1), 1–48 (2015)

Beck, N., Katz, J.N.: What to do (and not to do) with time-series cross-section data. Am. Polit. Sci. Rev. **89**(3), 634–647 (1995)

Beck, N., Katz, J.N.: Random coefficient models for time-series-cross-section data: Monte Carlo experiments. Polit. Anal. **15**(2), 182–195 (2007)

Bell, A., Jones, K.: Explaining fixed effects: random effects modelling of time-series cross-sectional and panel data. Polit. Sci. Res. Methods **3**(1), 133–153 (2015)
Bell, A., Johnston, R., Jones, K.: Stylised fact or situated messiness? The diverse effects of increasing debt on national economic growth. J. Econ. Geogr. **15**(2), 449–472 (2015)
Bell, A., Jones, K., Fairbrother, M.: Understanding and misunderstanding group mean centering: a commentary on Kelley et al'.s dangerous practice. Qual. Quant. **52**(5), 2031–2036 (2018)
Bell, A., Holman, D., Jones, K.: Using shrinkage in multilevel models to understand intersectionality: a simulation study and a guide for best practice (2018) **(in review)**
Blakely, T.A., Woodward, A.J.: Ecological effects in multi-level studies. J. Epidemiol. Community Health **54**(5), 367–374 (2000)
Brumback, B.A., Dailey, A.B., Brumback, L.C., Livingston, M.D., He, Z.: Adjusting for confounding by cluster using generalized linear mixed models. Stat. Probab. Lett. **80**(21–22), 1650–1654 (2010)
Brumback, B.A., Zheng, H.W., Dailey, A.B.: Adjusting for confounding by neighborhood using generalized linear mixed models and complex survey data. Stat. Med. **32**(8), 1313–1324 (2013)
Bryan, M.L., Jenkins, S.P.: Multilevel modelling of country effects: a cautionary tale. Eur. Sociol. Rev. **32**(1), 3–22 (2016)
Bullen, N., Jones, K., Duncan, C.: Modelling complexity: analysing between-individual and between-place variation—a multilevel tutorial. Environ. Plann. A **29**(4), 585–609 (1997)
Chatelain, J.-B., Ralf, K.: Inference on time-invariant variables using panel data: a pre-test estimator with an application to the returns to schooling. PSE Working Paper. https://ideas.repec.org/p/hal/wpaper/halshs-01719835.html (2018). Accessed 24 Apr 2018
Christmann, P.: Economic performance, quality of democracy and satisfaction with democracy. Electoral. Stud. **53**, 79–89 (2018). https://doi.org/10.1016/J.ELECTSTUD.2018.04.004
Clark, T.S., Linzer, D.A.: Should I use fixed or random effects? Polit. Sci. Res. Methods **3**(2), 399–408 (2015)
Croissant, Y., Millo, G.: Panel data econometrics in R: the plm package. J. Stat. Softw. **27**(2), 1–43 (2008)
Deeming, C., Jones, K.: Investigating the macro determinants of self-rated health and well-being using the European social survey: methodological innovations across countries and time. Int. J. Sociol. **45**(4), 256–285 (2015)
Delgado-Rodríguez, M., Llorca, J.: Bias. J. Epidemiol. Community Health **58**(8), 635–641 (2004)
Duncan, C., Jones, K., Moon, G.: Health-related behaviour in context: a multilevel modelling approach. Soc. Sci. Med. **42**(6), 817–830 (1996)
Fairbrother, M.: Two multilevel modeling techniques for analyzing comparative longitudinal survey datasets. Polit. Sci. Res. Methods **2**(1), 119–140 (2014)
Fairbrother, M.: Trust and public support for environmental protection in diverse national contexts. Sociol. Sci. **3**, 359–382 (2016). https://doi.org/10.15195/v3.a17
Fielding, A.: The role of the Hausman test and whether higher level effects should be treated as random or fixed. Multilevel Model. Newsl. **16**(2), 3–9 (2004)
Fotouhi, A.R.: Comparisons of estimation procedures for nonlinear multilevel models. J. Stat. Softw. **8**(9), 1–39 (2003)
Gelman, A.: Red State, Blue State, Rich State, Poor State : Why Americans Vote the Way They Do. Princeton University Press, Princeton (2008)
Gelman, A.: Why I don't use the term "fixed and random effects". Stat. Model. Causal Inference Soc. Sci. http://andrewgelman.com/2005/01/25/why_i_dont_use/ (2005). Accessed 19 Nov 2015
Goetgeluk, S., Vansteelandt, S.: Conditional generalized estimating equations for the analysis of clustered and longitudinal data. Biometrics **64**(3), 772–780 (2008)
Goldstein, H.: Multilevel Statistical Models, 4th edn. Wiley, Chichester (2010)
Greene, W.H.: Econometric Analysis, 7th edn. Pearson, Harlow (2012)
Grilli, L., Rampichini, C.: The role of sample cluster means in multilevel models: a view on endogeneity and measurement error issues. Methodology **7**(4), 121–133 (2011)
Grilli, L., Rampichini, C.: Specification of random effects in multilevel models: a review. Qual. Quant. **49**(3), 967–976 (2015)
Halaby, C.N.: Panel models in sociological research: theory into practice. Ann. Rev. Sociol. **30**(1), 507–544 (2004)
Hanchane, S., Mostafa, T.: Solving endogeneity problems in multilevel estimation: an example using education production functions. J. Appl. Stat. **39**(5), 1101–1114 (2012)
Hausman, J.A.: Specification tests in econometrics. Econometrica **46**(6), 1251–1271 (1978)
Heagerty, P.J., Kurland, B.F.: Misspecified maximum likelihood estimates and generalised linear mixed models. Biometrika **88**(4), 973–985 (2001)

Hedeker, D., Mermelstein, R.J.: Mixed-effects regression models with heterogeneous variance: Analyzing ecological momentary assessment (EMA) data of smoking. In: Little, T.D., Bovaird, J.A., Card, N.A. (eds.) Modeling Contextual Effects in Longitudinal Studies. Erlbaum, Mahwah, NJ (2007)
Herndon, T., Ash, M., Pollin, R.: Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. Camb. J. Econ. **38**(2), 257–279 (2014)
Howard, A.L.: Leveraging time-varying covariates to test within- and between-person effects and interactions in the multilevel linear model. Emerg. Adulthood **3**(6), 400–412 (2015)
Jones, K., Bullen, N.: Contextual models of urban house prices—a comparison of fixed-coefficient and random-coefficient models developed by expansion. Econ. Geogr. **70**(3), 252–272 (1994)
Jones, K., Subramanian, S.V.: Developing Multilevel Models for Analysing Contextuality, Heterogeneity and Change, vol. 2. University of Bristol, Bristol (2013)
Jones, K., Johnston, R., Manley, D., Owen, D., Charlton, C.: Ethnic residential segregation: a multilevel, multigroup, multiscale approach exemplified by London in 2011. Demography **52**(6), 1995–2019 (2015)
King, G., Roberts, M.: How robust standard errors expose methodological problems they do not fix. Polit. Anal. **23**(2), 159–179 (2015)
Kloosterman, R., Notten, N., Tolsma, J., Kraaykamp, G.: The effects of parental reading socialization and early school involvement on children's academic performance: a panel study of primary school pupils in the Netherlands. Eur. Sociol. Rev. **27**(3), 291–306 (2010)
Lauen, D.L., Gaddis, S.M.: Exposure to classroom poverty and test score achievement: contextual effects or selection? Am. J. Sociol. **118**(4), 943–979 (2013)
Leckie, G., Charlton, C.: runmlwin: a program to run the MLwiN multilevel modelling software from within Stata. J. Stat. Softw. **52**(11), 1–40 (2013). https://doi.org/10.18637/jss.v052.i11
Maas, C.J.M., Hox, J.J.: Robustness issues in multilevel regression analysis. Stat. Neerl. **58**(2), 127–137 (2004)
Maimon, D., Kuhl, D.C.: Social control and youth suicidality: situating durkheim's ideas in a multilevel framework. Am. Sociol. Rev. **73**(6), 921–943 (2008)
Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., Bates, D.M.: Balancing type I error and power in linear mixed models. J. Mem. Lang. **94**, 305–315 (2017)
McCulloch, C.E., Neuhaus, J.M.: Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. Stat. Sci. **26**(3), 388–402 (2011a)
McCulloch, C.E., Neuhaus, J.M.: Prediction of random effects in linear and generalized linear models under model misspecification. Biometrics **67**(1), 270–279 (2011b)
McNeish, D.: Small sample methods for multilevel modeling: a colloquial elucidation of REML and the Kenward–Roger correction. Multivar. Behav. Res. **52**(5), 661–670 (2017)
Milner, H.V., Kubota, K.: Why the move to free trade? Democracy and trade policy in the developing countries. Int. Org. **59**(1), 107–143 (2005)
Mundlak, Y.: Pooling of time-series and cross-section data. Econometrica **46**(1), 69–85 (1978)
Nerlove, M.: Essays in Panel Data Econometrics. Cambridge University Press, Cambridge (2005)
Neuhaus, J.M., McCulloch, C.E.: Separating between- and within-cluster covariate effects by using conditional and partitioning methods. J. R. Stat. Soc. Ser. B Stat. Methodol. **68**, 859–872 (2006)
Rasbash, J.: Module 4: multilevel structures and classifications. LEMMA VLE. http://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/4-concepts-sample.pdf (2008). Accessed 19 Nov 2015
Raudenbush, S.W., Bloom, H.S.: Learning about and from a distribution of program impacts using multisite trials. Am. J. Eval. **36**(4), 475–499 (2015)
Raudenbush, S.W., Bryk, A.: Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd edn. Sage, London (2002)
Raudenbush, S.W., Willms, J.: The estimation of school effects. J. Educ. Behav. Stat. **20**(4), 307–335 (1995)
Reinhart, C.M., Rogoff, K.S.: Growth in a time of Debt. Am. Econ. Rev. **100**(2), 573–578 (2010)
Ruiter, S., van Tubergen, F.: Religious attendance in cross-national perspective: a multilevel analysis of 60 countries. Am. J. Sociol. **115**(3), 863–895 (2009)
Sampson, R.J., Raudenbush, S.W., Earls, F.: Neighborhoods and violent crime: a multilevel study of collective efficacy. Science **277**(5328), 918–924 (1997)
Schempf, A.H., Kaufman, J.S., Messer, L., Mendola, P.: The neighborhood contribution to black-white perinatal disparities: an example from two north Carolina counties, 1999–2001. Am. J. Epidemiol. **174**(6), 744–752 (2011)
Schmidt-Catran, A.W.: Economic inequality and public demand for redistribution: combining cross-sectional and longitudinal evidence. Socio Econ. Rev. **14**(1), 119–140 (2016)

Schmidt-Catran, A.W., Fairbrother, M.: The random effects in multilevel models: getting them wrong and getting them right. Eur. Sociol. Rev. **32**(1), 23–38 (2015)
Schmidt-Catran, A.W., Spies, D.C.: Immigration and welfare support in germany. Am. Sociol. Rev. (2016). https://doi.org/10.1177/0003122416633140
Schurer, S., Yong, J.: Personality, well-being and the marginal utility of income: what can we learn from random coefficient models? Working Paper. https://ideas.repec.org/p/yor/hectdg/12-01.html (2012). Accessed 28 Apr 2018
Shin, Y., Raudenbush, S.W.: A latent cluster-mean approach to the contextual effects model with missing data. J. Educ. Behav. Stat. **35**(1), 26–53 (2010)
Shor, B., Bafumi, J., Keele, L., Park, D.: A Bayesian multilevel modeling approach to time-series cross-sectional data. Polit. Anal. **15**(2), 165–181 (2007)
Snijders, T.A.B., Bosker, R.J.: Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling, 2nd edn. Sage, London (2012)
Spiegelhalter, D.J.: Incorporating Bayesian ideas into health-care evaluation. Stat. Sci. **19**(1), 156–174 (2004)
Steele, F., Vignoles, A., Jenkins, A.: The effect of school resources on pupil attainment: a multilevel simultaneous equation modelling approach. J. R. Stat. Soc. Ser. A Stat. Soc. **170**, 801–824 (2007)
Stegmueller, D.: How many countries do you need for multilevel modeling? A comparison of frequentist and Bayesian approaches. Am. J. Polit. Sci. **57**(3), 748–761 (2013)
Subramanian, S.V., O'Malley, A.J.: Modeling neighborhood effects the futility of comparing mixed and marginal approaches. Epidemiology **21**(4), 475–478 (2010)
Subramanian, S.V., Jones, K., Kaddour, A., Krieger, N.: Revisiting Robinson: the perils of individualistic and ecologic fallacy. Int. J. Epidemiol. **38**(2), 342–360 (2009)
Vaisey, S., Miles, A.: What you can—and can't—do with three-wave panel data. Sociol. Methods Res. **46**(1), 44–67 (2017)
Vallejo, G., Fernández, P., Cuesta, M., Livacic-Rojas, P.E.: Effects of modeling the heterogeneity on inferences drawn from multilevel designs. Multivar. Behav. Res. **50**(1), 75–90 (2015)
Western, B.: Causal heterogeneity in comparative research: a bayesian hierarchical modelling approach. Am. J. Polit. Sci. **42**(4), 1233–1259 (1998)
Wooldridge, J.M.: Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge, MA (2002)
Zhang, Z., Parker, R.M.A., Charlton, C.M.J., Leckie, G., Browne, W.J.: R2MLwiN: a package to run MLwiN from within R. J. Stat. Softw. **72**(10), 1–43 (2016)