



Schmidt-Catran, A., & Fairbrother, M. H. (2016). The Random Effects in Multilevel Models: Getting Them Wrong and Getting Them Right. *European Sociological Review*, 32(1), 23-38.
<https://doi.org/10.1093/esr/jcv090>

Peer reviewed version

Link to published version (if available):
[10.1093/esr/jcv090](https://doi.org/10.1093/esr/jcv090)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This article has been accepted for publication in *European Sociological Review* Published by Oxford University Press

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

The Random Effects in Multilevel Models: Getting Them Wrong and Getting Them Right

Alexander Schmidt-Catran
University of Cologne

Malcolm Fairbrother
University of Bristol

4 January 2015

Abstract

Many surveys of respondents from multiple countries or sub-national regions have now been fielded on multiple occasions. Social scientists are regularly using multilevel models to analyze the data generated by such surveys, investigating variation across both space and time. We show, however, that such models are usually specified erroneously. They typically omit one or more relevant random effects, thereby ignoring important clustering in the data, which leads to downward biases in the standard errors. These biases occur even if the fixed effects are specified correctly; if the fixed effects are incorrect, erroneous specification of the random effects worsens biases in the coefficients. We illustrate these problems using Monte Carlo simulations and two empirical examples. Our recommendation to researchers fitting multilevel models to comparative longitudinal survey data is to include random effects at all potentially relevant levels, thereby avoiding any mismatch between the random and fixed parts of their models.

Introduction

Since 2001, the *European Sociological Review* has published 17 papers fitting multilevel models to comparative longitudinal survey data—observations on survey respondents collected in multiple countries or other contexts, with these higher-level units each observed multiple times. Of these papers, 10 reported models fitted with random effect (RE) structures we show below to be erroneous; using the correct structure might well have changed results central to many of these papers' main arguments and conclusions.

The incorrect specification of the REs in such models is an important concern, given that multilevel models have become a very prominent tool in the social sciences.¹ Used correctly, these models account for the contexts in which units of analysis are observed, a very useful property given that most social science theory speaks to the place of individuals in their social environments. More technically, compared with traditional single-level models, multilevel models account for dependencies in the data, thereby producing more appropriate inferences.²

Due to the increasing availability of comparative surveys collected over multiple waves (e.g. International Social Survey Programme, European Social Survey (ESS), World Values Survey), scholars have started to fit multilevel models not only to data from multiple countries but also to data that have been sampled at various time-points—survey data that are both comparative and, at the country level, longitudinal. While the nesting structure is obvious in the case of cross-sectional data (individuals observed at level-1 and countries at level-2), the correct structure is less obvious where countries are each observed multiple times. In this paper we demonstrate that the choice of the nesting structure in analyses of such data has important consequences for the substantive inferences typically derived from such models. We discuss six different modeling approaches and show that those used in a majority of published studies are problematic. Specifically, common errors in the specification of the REs lead to downwards-biased standard errors (SEs) for fixed effects (FEs).

Furthermore, if coefficients are biased due to misspecification of the FEs (e.g. an omitted variable), errors in the RE specification make such biases worse.

We show that the specification of the random part of a multilevel model must at least reflect the types of variables included in the fixed part. Failure to include REs at any level at which there are FEs can result in severely downwards-biased SEs and artificially inflated degrees of freedom for the coefficients on those FEs.³ In sum, the omission of relevant REs can severely increase the risk of making Type 1 errors. The two most common omissions in multilevel models of comparative longitudinal survey data are REs at the country and country-year levels. Omitting REs at the former level implies that respondents from Peru in 1990 and Peru in 1995 have no more in common, on average, than those from Peru in 1990 and Hungary in 1995. Omitting REs at the latter level implies that a respondent from Peru in 1990 has no more in common with another from Peru in 1990, on average, as with one from Peru in 1995. We believe these assumptions are unjustified.

The paper is organized as follows: We first discuss six possible RE structures, identifying examples of published research using such structures. In the next section, we report the results of a simulation study which assesses the performance of the six specifications in the context of data with known characteristics. Next we present two empirical applications—one a replication of a published paper, and the other an original analysis. In the final section, we conclude with the relatively simple recommendation that researchers should include RE at all relevant levels. The costs of omitting a necessary RE substantially outweigh the costs of including a redundant RE, and the theoretically possible situations where there is any cost at all of including a redundant RE are few and unlikely to occur in practice.

For ease of presentation, we refer throughout the paper to countries as the contextual-level units. But the arguments apply equally to data collected at multiple time-points from different regions nested within single countries (cities, provinces, counties, etc.), or to organizations such as schools or firms, as the examples to which we refer make clear.

Six Types of Models and the Relation between Fixed and Random Effects

If the data to be analyzed are drawn from multiple countries each observed on multiple occasions, a variety of potential RE specifications present themselves. REs can be included for countries, years, and combinations of countries and years (country-years). The range of possible model types, all but one of which has been used in existing studies based, is presented in Table 1.⁴

[Table 1 about here]

Models for comparative longitudinal survey data can incorporate covariates at each of these levels. Such covariates can be time-invariant country characteristics (e.g. common law legal tradition); characteristics which vary over time within countries (e.g. unemployment rate); or characteristics of years relevant for all countries (e.g. number of terrorist incidents globally). This final category of variables is likely to be relevant in few analyses of comparative longitudinal survey data; to the best of our knowledge, no cross-national studies have used such a year-level variable. However, among analyses of survey data drawn from multiple regions within a given country, there are some applications which estimate the effect of year-level variables; Schlueter and Davidov (2013) for example include a measure of national media content, in their analysis of perceived group threat. More typically, models of comparative longitudinal survey data will include variables only at the country and country-year levels.

To explain the implications of the six models presented in Table 1, we refer to a hypothetical dataset comprising respondents from 20 countries surveyed in 5 different years.

Model A has two levels and considers respondents to be nested within 100 country-years. This structure implicitly treats any time-invariant country-level variable as a time-varying one, and thus as having 100 unique values. Conversely, the model does not regard repeated observations on countries as nested within countries. Papers employing such a structure include Huijts *et al.* (2010) and Semyonov *et al.* (2006). We replicate the former study below; the latter study, of anti-immigrant

attitudes, rests on models of respondents nested in 12 countries each observed four times—which the authors describe as “48 (12 × 4) observations at the country level” (p. 439).

Model B also has two levels, but this specification emphasizes the nesting of individuals within the 20 countries. For this model, any variable that is non-constant within a country—such as a national level condition that varies over time but is constant for all respondents observed in a given country-year—will be taken as a property of individuals, since there is no clustering below the country level. Eger (2010) uses such a model in studying support for the welfare state across Sweden’s 21 counties, where each county is observed four times. Several covariates she labels as county-level are non-constant within counties.

Model C has three levels, with years taken as the highest level, and country-years nested within them. This model is similar to B, but recognizes that respondents within any given year—irrespective of where they are—have more in common than respondents from different years. The model would have 5 year clusters at the highest level and 100 country-year clusters at the level below. We found no applications of models like this with international survey data, but some based on repeated cross-sections within countries (Andersen *et al.*, 2006; Schlueter and Davidov, 2013). Andersen *et al.* (2006) examined British voters nested in 571 constituencies over eight elections, using models that “account for the clustering of individuals within constituencies, and constituencies within years” (p. 218).

Model D, another three-level model, ignores the clustering of respondents from all places within years (unlike C), but (like B) recognizes that respondents from the same country are more similar than respondents from different countries. In addition to Model B it recognizes that respondents observed in the same country in the same year have more in common than respondents observed in the same country but in a different year. This specification assumes that the 100 country-years at level 2 are nested in 20 countries at level 3. Fairbrother (2013) has used a specification of this type in studying the correlates of environmental concern.

Model E is a cross-classified model with both 20 countries and 5 years as the higher-level units. In contrast to Models C and D, this approach does not assume hierarchical nesting. It

recognizes that respondents observed within a given year are likely to be more similar than respondents from different years, and that respondents observed within a given country are likely to be more similar than respondents from different countries; but within a given year or country respondents are no more likely to be similar if they are observed in the same country or year, respectively. Lubbers and Scheepers (2001) have made use of cross-classified models of this type, in analyzing voting for the extreme right across 17 regions in Germany.

Model F, finally, includes REs for (20) countries, (5) years, and (100) country-years. Model F treats country-years as cross-classified within countries and years, and individuals as strictly nested in one higher-level unit: country-years. We are not aware of any studies which have applied a model of type F to comparative longitudinal survey data.

Our presentation of this typology of models has not been innocent; the limitations of some RE structures should already be clear. Models A and C assume too much independence in taking the number of distinct values of country-level variables as 100 when it is actually 20, or alternatively in assuming independence of the 5 repeated observations of time-variant country-level variables. Structures B and E suffer from a different problem, if they include country-year-level variables (as did both studies we referenced). They do not take into account that country-year-level variables are themselves clustered; because country-year-level variables are not constant within countries, they are implicitly taken as individual-level variables, inflating the degrees of freedom and deflating the SEs. A cluster-level variable is defined by the fact that it is constant within clusters; a variable that is non-constant within clusters cannot by definition describe those clusters.

Structures A and B, and also D, suffer from the limitation that they ignore a potentially relevant level: years. If there is random variation between years the model will yield biased SEs for any year-level variables. Finally, Model F is a full model, which should in principle control for any possible statistical dependence and includes a level for any kind of variable. As Model F is the most complex, it might face problems of convergence in cases where Model D is feasible, since many applications will include no year-level variables, making Model D's one limitation inconsequential.

A Review of Published Research

In order to assess the prevalence of each of these specifications in applied research, we identified all relevant articles published in the *European Sociological Review* (ESR) since 1985 and coded them according to our typology. We searched for the term “multilevel” in the journal’s online search engine, which returned 191 articles.⁵ In the first stage, we selected all those characterized by a multilevel structure, with individuals nested in contextual units observed on multiple occasions, and which estimate at least one contextual-level effect. This resulted in a set of 34 articles that could potentially use one of the model types we presented above. In the next stage, we excluded articles that used individual-level panel data (N=14) because the appropriate RE structure in these models would be even more complex—entailing an additional level of clustering.⁶ We furthermore excluded two papers for which we were not able to determine the type(s) of variables, and we excluded one paper which presented a model without any RE, thus falling outside our typology. This left us with a sample of 17 papers which fitted multilevel models to comparative longitudinal survey data and included at least one contextual-level variable. Table 2 presents these papers.

[Table 2 about here]

We identified three papers using Model A, and 11 papers of type B. Six of the latter included only time-invariant country-level variables—which were in every case averages of time-varying variables. These models could have exploited variation over time in addition to variation across space, an issue we discuss further below. We also found two papers that used Model C, and one that used Model D. The results of our keyword search did not turn up any articles in *ESR* presenting models of type E. However, separately we identified one such paper, which we discussed earlier (Lubbers and Scheepers, 2001). A majority of the papers included analyses using RE structures which we show in the next section led to downwards-biased SE. We should note that the authors of these papers may have tested other RE specifications and not presented them, because the results were identical. In this sense, we do not know that the results from the reviewed papers are erroneous; but the results as presented, in the absence of results from fuller models, are highly questionable.

A Simulation Study

To investigate the consequences of using different RE structures, we now present a simulation study. We employ 14 data-generating processes (DGPs). All of them include covariates at each of four levels (individual, country-year, country, year). Except where specified otherwise, all covariates are drawn from a normal distribution (with a mean of 0 and standard deviation of 1) and are uncorrelated with all REs. The residual error has a mean of 0 and a variance of 4. In the first eight DGPs each simulated dataset comprises 10 respondents observed in each of 5 years in each of 20 countries. In these DGPs all covariates are uncorrelated with each other and the estimated models are correctly specified given the fixed part in the DGPs. What varies across these eight DGPs are the variances of the RE at each of the contextual levels, for the purpose of testing the implications of omitting a random intercept at a level with random error. Next, in DGPs 9, 10 and 11, we vary the N s at the individual and year level. In DGPs 12, 13 and 14, we analyze the impact of misspecifications in the fixed part. In 12 and 14, we simulate data with correlated covariates, omitting one in the estimation. In 13 and 14 the cross-sectional and longitudinal effects of a time-varying country-level covariate differ. The full R code used for the simulations, and for generating the graphics presented below, is available from the authors on request. To ensure the robustness of the results of our simulation study, we ran two separate series of 10,000 simulations (starting with different random number seeds), inspected the chains visually, and verified that the results were substantially the same.

The N s we have chosen at some levels are not realistic, but keeping them modest allowed us to run more simulations, and below we briefly report the results of some tests of how our results varied depending on the N at each level. Particularly given that some models have an N of five (years) at the highest level—too few for the reliable estimation of a random effect variance—we experimented with the use of dummy variables rather than random intercepts for years; this alternative approach yielded otherwise identical results. Note that Model D with the addition of year dummies is the same as Model E with year dummies in place of random effects for years.

[Table 3 about here]

Consequences of Misspecifications in the Random Part

In this section we discuss the first 11 DGPs, which can be written as follows:

$$y_{ijkl} = 1 \cdot X_{ijkl} + 1 \cdot Z_{jkl} + 1 \cdot Z_l + 1 \cdot Q_k + u_{jkl} + u_l + u_k + e_{ijkl}$$

The subscript i indicates individuals, j indicates country-years, l indicates countries and k indicates years. X_{ijkl} is an individual-level and Q_k a year-level variable. Z varies at two levels, the country and the country-year level, with Z_l the between-country component and Z_{jkl} the within-country component. In DGPs 1-12 we set these two components to have the same effect.

To 10,000 datasets of each type, we fit two sets of six models. Each set of six includes the six RE structures discussed above. The first set includes separate country-year and country-level covariates, while the second set forces these two covariates to share a single coefficient—as did all but one of the models in the applied work we discussed earlier. The fixed part of these models replaces the distinct within and between components with a single variable (the estimated *fixed part* is $y_{ijkl} = \beta_0 + \beta_1 X_{ijkl} + \beta_2 Z_{2,jkl} + \beta_3 Q_k$, where $Z_{2,jkl} = Z_l + Z_{jkl}$). Studies treating time-varying national characteristics like GDP/capita as country-year-level variables implicitly assume that cross-sectional differences among countries have the same relationship with y as longitudinal changes within them. In the DGPs 1-12 this assumption is true and therefore both sets of six yield unbiased coefficient estimates.

Figure 1 presents the first key results of the simulation study: the implications of different RE structures for the SEs of FEs at different levels, across DGPs 1-8.

[Figure 1 about here]

The optimism of the SEs is defined as the ratio of a coefficient's sampling variability across the 10000 simulations to the variability as estimated by the SEs (Shor *et al.*, 2007). In Figure 1 we display the optimism on a log-scale, meaning that values greater than 0 indicate downwards-biased and anticonservative SEs and values smaller than 0 indicate upwards-biased, overly conservative SEs.

Figure 1 shows that the SEs associated with country-year-level variables are too small where a fitted model excludes RE at the country-year level (B and E), assuming there is variance at that level

(DGPs 1, 3, 5, 7). Even if there is no random error at the country-year level, if there is random error at the year level (DGPs 2, 6), the SEs for country-year-level covariates are too small using Model B, which excludes country-year REs. The country-year level in Model B reflects year-level error that is not allowed for by the RE structure. Whenever there are random differences across country-years—arising from random variance at the country-year and/or year levels—the SEs are biased at this level.

For year-level covariates, the SEs are too small if there is random error at the year level and the fitted model excludes REs at this level (DGPs 1, 2, 5, 6; Models A, B, D). Country-level covariates have downwards-biased SEs when there is random error at the country level (DGPs 1, 2, 3, 4) but the model does not include REs at that level (A, C). Finally, where a time-varying country-level variable is included but not decomposed into its country- and country-year component (the bottom row), the bias in SEs is a combination of the biases found for each of the components.

Overall, then, only Model F avoids the problem of anticonservative SEs for all FEs, no matter what random error is represented in the data. Model D also avoids the problem of anticonservative SEs for all coefficients, except those on year-level variables, since it includes no year-level RE. On the other hand, the SEs for some covariates as estimated by Model F and even more so Model D are, in a small number of exceptional circumstances, over-conservative. Where a random effect is estimated at a level where there is in fact no random error at all, the log-optimism of the SE for a covariate at that level can fall below 0 (see e.g. the results for DGPs 5 and 6 under Models A and D in the top row in Figure 1). Nevertheless, the overconservatism for all SEs as estimated by Model F is small, and in any event it seems unlikely in the extreme for there to be precisely zero random error at any level in real-world data.

DGPs 9-11 vary the numbers of years and individuals. The results (not presented) show that the numbers of years and individuals have no influence in correctly specified models (F, and D if one is concerned only with country- and country-year-level variables). In those models that take country-year level variables as individual-level variables (B and E), an increase in the number of individuals and an increase in the number of years both make the bias in SE worse. In models that treat

repeated observations of time-constant country-level variables as distinct values (A and C), the bias in SE increases with the number of years.

Consequences of Misspecifications in the Random and Fixed Part

Now we present the simulation results for DGPs 12-14, in which we investigate consequences of misspecifications in the random *and* fixed parts of a model. All models fitted to DGPs 1-11 had correctly specified fixed parts. However, for each DGP we fitted two sets of models, one in which we estimated separate country- and country-year-level effects and one in which we forced them to share a coefficient, like all but one of the published papers we discussed earlier. This second set yielded unbiased coefficients because the two components were created with identical effects. All the FE coefficients for DGPs 1-8 are unbiased (results not presented).

Typically such an assumption makes little sense in light of the usual results of Hausman (1978) tests, which regularly show that cross-sectional and longitudinal relationships are not equivalent (Halaby, 2004: 527). Fortunately, such a misspecification is easy to avoid, the solution being to include the country means and the mean-differenced components of time-varying country-level variables in the model. Fairbrother (2014) provides a detailed treatment of such models in the context of comparative longitudinal survey data and we invite readers to look at this paper for more details. Not centering time-varying country-level variables by their country-means generates a single coefficient representing a weighted mean of the between and within effects (Raudenbush and Bryk, 2002: 134-49). This coefficient can be deeply misleading and/or difficult to interpret if the two effects are different. Here we investigate what happens if the between and within effects differ, by setting the effect size of Z_{jkl} to 0.5 instead of 1 in DGP 13.

To investigate the consequences of another common fixed part misspecification, the omission of a relevant variable, in DGP 12 we add two covariates (Z_{jkl}^{CC} and Z_l^{CC}):

$$y_{ijkl} = 1 \cdot X_{ijkl} + 1 \cdot Z_{jkl} + 1 \cdot Z_l + 1 \cdot Z_{jkl}^{CC} + 1 \cdot Z_l^{CC} + 1 \cdot Q_k + u_{jkl} + u_l + u_k + e_{ijkl}, \quad \text{where}$$

$\text{Corr}(Z_l, Z_l^{CC}) = 0.8$. This is, we add a second country-year-level covariate, whose country-level

component is correlated with the first covariate (thus the superscript CC). Finally, in DGP 14 we combine these two characteristics by including the correlated covariate, as in DGP 12, and setting the effect of Z_{jkl} to 0.5, as in DGP 13. To these three DGPs we fit two sets of six models; the first set of six controls for the new covariate, while the second set of six omits it from the model. Both sets of six do not estimate separate country- and country-year-level effects but only the combined effect of Z. Figure 2 presents the mean FE coefficient across the 10000 simulations.

[Figure 2 about here]

The coefficient on the combined $Z_{2,jkl}$ is unbiased in DGP 12 if the correlated covariate is controlled for but biased if it is omitted from the model, and the severity of the bias depends on which of the six models is fitted. The bias is most severe for Models A and C, where the RE specification excludes the level at which the variables are correlated (the country level). In DGP 13, similarly, the coefficient estimate depends on which of the six models is estimated. In Models A and C, which do not recognize that there are fewer countries than country-years, the within and between components receive the same weight, which results in an estimated effect that is the average of the two components. In Models that take the within-country component Z_{jkl} as an individual-level variable (B and E), the estimated effect for the combined variable $Z_{2,jkl}$ is heavily dominated by this within effect. In those models that do recognize the correct number of countries and country-years (D and E), the effect is slightly dominated by the within component. These results are in line with Raudenbush and Bryk (2002) who noted that the combined effects are a weighted average of the within and between components, where the weights depend (approximately) on the degrees of freedom at each level.

The results for DGP 14, which combines the key features of DGPs 12 and 13, shows most dramatically of all how changing the RE structure alone can lead to a variety of different coefficient estimates, when the fixed part is misspecified. Here the estimated coefficient does not just vary, but can be either positive or negative, depending on the RE specification.

The Empirical Relevance of Misspecifications

Here we further illustrate the consequence of misspecified REs by first replicating a published study whose results change if we specify a different RE structure, and then presenting an original empirical example. In the latter case, fitting all six models demonstrates how much results can change depending on the RE structure.

A Replication Study

We contacted the authors of several papers to ask for replication datasets. Though some in principle said they were willing, ultimately none provided us with access to their data. In some cases, this reticence was understandable: the authors committed to uphold data protection protocols. In other instances, the authors provided no clear explanation why they were unwilling to provide replication data.

We therefore replicate one paper published in the *ESR* whose data were publicly available, and so whose authors we did not need to contact: “Education, Educational Heterogamy, and Self-Assessed Health in Europe: A Multilevel Study of Spousal Effects in 29 European Countries” by Huijts *et al.* (2010). The authors provided the contextual-level data in the paper, which was a commendably transparent approach on their part. Their analysis is based on the first three rounds of the ESS.

The dependent variable is individual subjective health. The main interest is in the effect of educational heterogamy, a contextual-level variable that measures the share of couples with the same educational level. This variable is measured at the country-year level, i.e. it varies between waves. As controls the model includes one country-level variable (government health expenditures as percentage of all health expenditures) and a second country-year-level variable (logged GDP/capita in 1,000 US\$). The authors estimate Model A, treating each country-year as a single observation in a two-level model.

We are not able to replicate precisely the results because the ESS version to which we have access differs from the version that the authors had.⁷ Our analysis is based on three country-years

less than the original study (Sweden 2002 and 2004, Iceland 2004). However, the estimates of our replication are sufficiently close to the original estimates to make our point. Table 4 presents the results of the original study (Huijts *et al.*, 2010: 270), our replication and our re-estimation of the model with the nesting structure of Model D.

[Table 4 about here]

In the original study M0 is an empty model and M1 a model containing all individual-level variables. In Model M2 the authors enter the degree of educational heterogamy (see Table 4). It has an effect of .011 and is significant at the 5% level. In our replication we estimate a coefficient of .011 but with a slightly higher SE such that the effect is not significant; nevertheless, we are able to reproduce a nearly identical result. In the re-estimated model (termed Alternative), the effect becomes nowhere near significant, and the coefficient is dramatically different.

More interesting is Model M3 because it also contains a country-level variable (government health expenditure). The authors conclude from Model M3 that “government share in health expenditure ... is negatively related to self-assessed health: the larger the financial role of the government in the health system, the less healthy people feel. This may indicate that health systems in which the government covers a large share of health expenses are less successful in improving and maintaining people’s health than health systems in which private funds play a more prominent role”(Huijts *et al.*, 2010: 270). This is clearly a far-reaching conclusion, and not an argument to be made lightly. Our replication of M3 with the inclusion of a country-level RE shows that this result is not robust. The absolute value of the coefficient on government health expenditure shrinks substantially, the SE expands slightly, and the effect turns non-significant. Is this estimate more defensible than the original? Given the simulations we reported earlier, it would seem that the use of Model A artificially expands the number of observations of country-level variables, downward-biasing the SE. In this application, there are 29 independently observed values but the model assumes the data contain 58. The re-estimated model takes the country level into account, and therefore reflects the correct number of independent observations. The estimated coefficient on

GDP/capita, a country-year-level variable, also changes under the new specification. The effect size declines substantially, though the coefficient remains significant.

The replication and re-estimation of the study therefore demonstrate the consequences of omitting relevant REs for the SEs of coefficients. Moreover, as demonstrated in the simulations, changes in the RE specification can have substantial impacts on the coefficient estimates. In the next sub-section we present an empirical example that demonstrates even more clearly how substantial these impacts can be.

An Empirical Example with Real Survey Data

The analyses presented in Table 5 are based on five waves of the ESS. The dependent variable is respondent's support for income redistribution. The survey data have been merged with four country-year-level variables: GDP/capita, Gini index, social spending as percentage of GDP, and the unemployment rate. The complete data set is taken from an article published elsewhere (*blinded for review*); readers should refer to this paper if they are interested in details about the data or the underlying theory. The main right-hand side variable of interest is income inequality (measured via the Gini index), with which there are theoretical reasons (derived from the median-voter hypothesis) to expect a positive correlation with demand for redistribution. GDP/capita, social spending and unemployment rates are included as control variables.

[Table 5 about here]

While the replication study presented above demonstrated how the FE estimates of Model A change if Model D is estimated instead, Table 5 shows how the FE estimates can vary depending on which of all six models are estimated. The coefficients on the individual-level variables change to a small degree, but not substantially. The coefficients and SE for the country-year-level variables, however, change substantially. The coefficient on GDP/capita is positive and significant in Model B (.0053***), while it turns significantly negative in Models C (-.0064*) and E (-.0042**). In Models A, D and F the effect is not significant. We observe a similar pattern for the effect of social spending: some models suggest a significant positive effect while others suggest a significant negative effect. The other two

contextual-level variables do not change quite so dramatically, but nonetheless vary and point to rather different conclusions. The effect of the Gini index is significantly positive in Models B, D, E and F but not significantly different from zero in Models A and C. The effect of unemployment rates is significantly positive in Models A and C but not significant in all other models. Model F (as well as model D) supports the conclusion that inequality is positively related to demand for redistribution while all other country-year-level variables have non-significant effects. For the purpose of our paper, these results all show how a model with the same FE specification can lead to substantially different conclusions, if the random part changes.

In our discussion of the six model types we pointed to the similarities of models A and C which both ignore country-level REs. These similarities are reflected in the coefficient estimates presented in Table 5. The simulation study showed that the SEs of country-year-level variables not centered by the country-mean are too optimistic, for multiple DGPs using all RE structures except D and F (see the bottom panel in Figure 1). This is also reflected in the estimates: While models D and F report only one significant effect, all other models show two or three significant effects. In an additional analysis (not presented here), we estimate the models from Table 5 but distinguish within- and between-country effects. This analysis shows that the between effect of, for example, the Gini index is close to 0 but the within effect is positive (about .02). Consequently, Models A and C in Table 5 estimate a Gini effect that lies between 0 and .02 while Models B, D, E and F provide estimates that are dominated by the within effect. Similar stories could be told about the other three variables. Generally, the additional analyses show that the changes in FE coefficients are much smaller than they were in the analysis without decomposition. There are no longer any instances where an effect changes substantially, which is analogous to the results of the simulation study, where correctly specified FEs are unaffected by the RE specification. The coefficients do still change slightly because—with this real survey data—there are certainly more misspecifications than just the missing decomposition into within- and between effects (e.g. omitted variables: see Figure 2, DGP 14).

Summary and Conclusion

All of the above points to a single specific, but important, practical implication: it is safer to include REs at all potentially relevant levels. The first and general rule is to include random effects at all levels at which there are fixed effects. A second rule however is: If there might be variation at a given level, even if there are no FEs included at that level, one should include REs at that level. This statement is particularly true for the level of geographical clusters (e.g. countries), as there will certainly be (unobserved) differences among these clusters that are stable over time. Observations on the same geographical clusters will very likely not be independent of each other, and so omitting REs at this level will lead to downwards-biased SEs.

Our coding of relevant articles published in *ESR* shows that several papers did not include country REs (Model A), even though they included country-level FEs (violating rule 1). Some papers included country-year-level FEs and REs but no country-level REs (Model C). In this case there is no mismatch between the random and the fixed part of the model, but the assumption of independence of the repeated observations of geographical clusters, as implied by this RE structure, might also lead to downwards-biased SE (violating rule 2). Moreover, many (though not all) of the models of type B listed in Table 2 included FEs at the country-year and/or year levels, but did not include country-year or year REs. These models again suffer from a mismatch between the fixed and random parts (violating rule 1). Some models of type B included only time-invariant country-level variables, for which the SE should be unbiased, though they did not take advantage of the longitudinal properties of the data.

For a very practical reason different applied researchers may be thinking about this issue in different ways: some software requires users to specify the level of a variable, and provides a warning when variables are not constant within the relevant clusters. Other software, however, does not. Mplus and HLM require specifying the measurement level of each variable and thereby force users to become aware of this issue. Mplus even aborts estimation if cluster-level variables are not constant within clusters. MLwiN and HLM provide model equations, with subscripts indicating at

which level the variables are assumed to be measured. In SPSS, Stata, R and SAS it is not necessary to specify the measurement levels. However, SPSS and SAS provide approximate degrees of freedom, which is indirect information about the measurement levels.

Excluding relevant REs can lead to severely misleading inferences. Our recommendation to researchers is thus similar to Barr *et al.*'s (2013), in their recent demonstration of the importance of random slopes in multilevel models: “keep it maximal”. If there is absolutely no random variation at a given level, our simulation study has shown that redundant random effects may yield over-conservative standard errors. But assuming that a given model converges, including a redundant random effect should in practice do no harm, as there will always be some random error at any level.

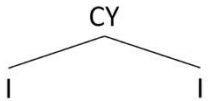
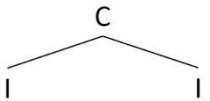
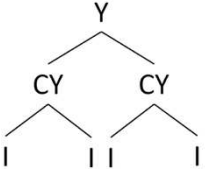
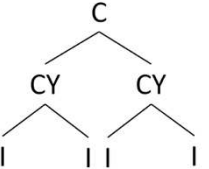
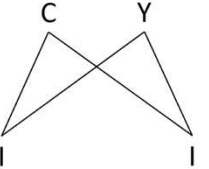
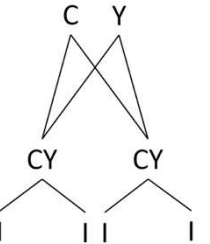
References

- Andersen, R., Yang, M. and Heath, A. F. (2006). Class Politics and Political Context in Britain, 1964–1997: Have Voters Become More Individualized?, *European Sociological Review*, **22**, 215-228.
- Barr, D.J., Levy, R., Scheepers, C. and Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal, *Journal of Memory and Language*, **68**, 255–278.
- Bell, A. and Jones, K. (2014). Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data, *Political Science Research and Method*, **3**, 133-153.
- Biedinger, N., Becker, B. and Rohling, I. (2008). Early Ethnic Educational Inequality: The Influence of Duration of Preschool Attendance and Social Composition, *European Sociological Review*, **24**, 243-256.
- Dinesen, P. T. (2013). Where You Come From or Where You Live? Examining the Cultural and Institutional Explanation of Generalized Trust Using Migration as a Natural Experiment, *European Sociological Review*, **29**, 114-128.
- Eger, M. A. (2010). Even in Sweden: The Effect of Immigration on Support for Welfare State Spending, *European Sociological Review*, **26**, 203-217.
- Engelhardt, H. (2012). Late Careers in Europe: Effects of Individual and Institutional Factors, *European Sociological Review*, **28**, 550-563.
- European Social Survey. Round 1-4 Cumulative Data (2011): *Data file Edition 1.0*; Round 5 Data (2012): *Data file Edition 2.0*. Norwegian Social Science Data Services, Norway.
- Fairbrother, M. (2013). Rich People, Poor People, and Environmental Concern: Evidence across Nations and Time, *European Sociological Review*, **29**, 910-922.
- Fairbrother, M. (2014). Two Multilevel Modeling Techniques for Analyzing Comparative Longitudinal Survey Datasets, *Political Science Research and Methods*, **2**, 119-140.

- Fekjær, S. N. and Birkelund, G. E. (2007). Does the Ethnic Composition of Upper Secondary Schools Influence Educational Achievement and Attainment? A Multilevel Analysis of the Norwegian Case, *European Sociological Review*, **23**, 309-323.
- Gerlitz, J.-Y., Mühleck, K., Scheller, P. and Schrenker, M. (2012). Justice Perception in Times of Transition: Trends in Germany, 1991–2006, *European Sociological Review*, **28**, 263-282.
- Halaby, C.N. (2004). Panel Models in Sociological Research: Theory into Practice, *Annual Review of Sociology*, **30**, 507–544.
- Hausman, J.A. (1978). Specification Tests in Econometrics, *Econometrica*, **46**, 1251–1271.
- Huijts, T., Monden, C. W. S. and Kraaykamp, G. (2010). Education, Educational Heterogamy, and Self-Assessed Health in Europe: A Multilevel Study of Spousal Effects in 29 European Countries, *European Sociological Review*, **26**, 261-276.
- Immerzeel, T. and van Tubergen, F. (2013). Religion as Reassurance? Testing the Insecurity Theory in 26 European Countries, *European Sociological Review*, **29**, 359-372.
- Kalmijn, M. (2010). Country Differences in the Effects of Divorce on Well-Being: The Role of Norms, Support, and Selectivity, *European Sociological Review*, **26**, 475-490.
- King, G. and Roberts, M.E. (2014). How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It, *Political Analysis* (online first): doi:10.1093/pan/mpu015.
- Kogan, I. and Kalter, F. (2006). The Effects of Relative Group Size on Occupational Outcomes: Turks and Ex-Yugoslavs in Austria, *European Sociological Review*, **22**, 35-48.
- Lubbers, M. and Scheepers, P. (2001). Explaining the Trend in Extreme Right-wing Voting: Germany 1989-1998, *European Sociological Review*, **17**, 431-449.

- Meulemann, H. (2004). Enforced Secularization — Spontaneous Revival?: Religious Belief, Unbelief, Uncertainty and Indifference in East and West European Countries 1991–1998, *European Sociological Review*, **20**, 47-61.
- Meulemann, H. (2012). Information and Entertainment in European Mass Media Systems: Preferences for and Uses of Television and Newspapers, *European Sociological Review*, **28**, 186-202.
- Moerbeek, M. (2004). The Consequence of Ignoring a Level of Nesting in Multilevel Analysis, *Multivariate Behavioral Research*, **39**, 129-149.
- Raudenbush, S.W. and Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd ed.* London: Sage.
- Schlueter, E. and Davidov, E. (2013). Contextual Sources of Perceived Group Threat: Negative Immigration-Related News Reports, Immigrant Group Size and their Interaction, Spain 1996–2007, *European Sociological Review*, **29**, 179-191.
- Semyonov, M., Raijman, R. and Gorodzeisky, A. (2006). The Rise of Anti-foreigner Sentiment in European Societies, 1988-2000, *American Sociological Review*, **71**, 426-449.
- Shi, Y., Leite, W. and Algina, J. (2010). The impact of omitting the interaction between crossed factors in cross-classified random effects modelling, *British Journal of Mathematical and Statistical Psychology*, **63**, 1-15.
- Shor, B., Bafumi, J., Keele, L. and Park, D. (2007). A Bayesian Multilevel Modeling Approach to Time-Series Cross-Sectional Data, *Political Analysis*, **15**, 165–181.
- Stegmueller, D., Scheepers, P., Roßteutscher, S. and de Jong, E. (2012). Support for Redistribution in Western Europe: Assessing the role of religion, *European Sociological Review*, **28**, 482-497.
- van der Lippe, T., de Ruijter, J., de Ruijter, E. and Raub, W. (2011). Persistent Inequalities in Time Use between Men and Women: A Detailed Look at the Influence of Economic Circumstances, Policies, and Culture, *European Sociological Review*, **27**, 164-179.

Table 1: A Typology of Random Effects Structures for Multilevel Models of Comparative Longitudinal Survey Data

Random Effects	Model A	Model B	Model C	Model D	Model E	Model F
Country		✓		✓	✓	✓
Year			✓		✓	✓
Country-Year	✓		✓	✓		✓
Structure						
						

Notes: C = Country-level RE, Y = Year-level RE, CY = Country-year-level RE, I = Individual-level.

Table 2: A Review of Articles from the *European Sociological Review*

ID	Fixed Effects			RE Structure			Type	Authors	Year
	Country	Year	Country-Year	Country	Year	Country-Year			
76	✓		✓			✓	A	Gerlitz <i>et al.</i>	2012
86	✓		✓			✓	A	van der Lippe <i>et al.</i>	2011
116	✓		✓			✓	A	Huijts <i>et al.</i>	2010
30	✓			✓			B	Immerzeel and van Tubergen	2013
44	✓			✓			B	Dinesen	2013
62	✓			✓			B	Stegmueller <i>et al.</i>	2012
67	✓		✓	✓			B	Meulemann	2012
78	✓			✓			B	Engelhardt	2012
114	✓			✓			B	Kalmijn	2010
118			✓	✓			B	Eger	2010
149	✓			✓			B	Biedinger <i>et al.</i>	2008
158			✓	✓			B	Fekjær and Birkelund	2007
166			✓	✓			B	Kogan and Kalter	2006
176	✓	✓	✓	✓			B	Meulemann	2004
17		✓	✓		✓	✓	C	Schlueter and Davidov	2013
173			✓		✓	✓	C	Andersen <i>et al.</i>	2006
40	✓		✓	✓	✓		D	Fairbrother	2013

Notes: Country level variables are time-invariant and describe the geographical higher-level units. Year-level variables vary only between years. Country-year level variables are time-varying and describe the geographical higher-level units. Model types correspond to Table 1. Full information on our coding rules is available from the authors upon request.

Table 3: Random Effects Variances and Number of Units for DGPs 1-14

DGP	N_i	N_k	N_i	$\text{Var}(u_i)$	$\text{Var}(u_k)$	$\text{Var}(u_{jkl})$	Correlated Coefficient	Effect of Z_{jkl}
1	20	5	10	1	1	1	N	1
2	20	5	10	1	1	0	N	1
3	20	5	10	1	0	1	N	1
4	20	5	10	1	0	0	N	1
5	20	5	10	0	1	1	N	1
6	20	5	10	0	1	0	N	1
7	20	5	10	0	0	1	N	1
8	20	5	10	0	0	0	N	1
9	20	5	50	1	1	1	N	1
10	20	2	10	1	1	1	N	1
11	20	20	10	1	1	1	N	1
12	20	5	10	1	1	1	Y	1
13	20	5	10	1	1	1	N	0.5
14	20	5	10	1	1	1	Y	0.5

Notes: index l indicates countries, k indicates years, i indicates individuals and j indicates country-years.

Table 4: Replication and Re-estimation of Huijts *et al.* (2010)

	M2						M3							
	Original b/se		Replication b/se		Alternative b/se		Original b/se		Replication b/se		Alternative b/se			
Constant	3.093	** 7	0.19	0.21	3.100	** 7	0.21	0.15	2.229	** 7	0.22	0.32	2.832	** 4
<i>Individual characteristics</i>														
Educational level														
Primary (ref.)														
Secondary	0.134	** 9	0.00	0.00	0.135	** 9	0.136	** 9	0.135	** 9	0.135	** 9	0.136	** 9
Tertiary	0.237	** 0	0.01	0.01	0.247	** 1	0.247	** 1	0.238	** 0	0.247	** 1	0.247	** 1
Educational level partner														
Primary (ref.)														
Secondary	0.076	** 9	0.00	0.00	0.087	** 9	0.087	** 9	0.076	** 9	0.087	** 9	0.087	** 9
Tertiary	0.112	** 0	0.01	0.01	0.129	** 1	0.130	** 1	0.112	** 0	0.130	** 1	0.129	** 1
Father's occupation														
Manual and service (ref.)														
Technical and craft	0.018	* 8	0.00	0.01	0.024	* 0	0.023	* 0	0.018	* 8	0.023	* 0	0.022	* 0
Clerical and intermediate	0.041	* 7	0.01	0.02	0.054	* 1	0.052	* 1	0.040	* 7	0.053	* 1	0.052	* 1
Professional	0.045	** 2	0.01	0.01	0.062	** 5	0.060	** 5	0.045	** 2	0.061	** 5	0.061	** 5
Managers and administrators	0.037	** 2	0.01	0.01	0.036	* 4	0.034	* 4	0.037	** 2	0.035	* 4	0.034	* 4
No known occupation	-	0.01	-	0.01	-	0.01	-	0.01	-	0.01	-	0.01	-	0.01
Gender (female=1)	0.081	** 6	0.00	0.00	0.080	** 7	0.080	** 7	0.081	** 6	0.080	** 7	0.080	** 7
Age	0.018	** 0	-	0.00	0.018	** 0	0.018	** 0	0.018	** 0	0.018	** 0	0.018	** 0
<i>Country characteristics</i>														
Educational heterogamy	0.011	* 0.00	0.011	0.00	0.003	0.00	-	0.00	-	0.00	-	0.00	-	0.00

			5		6		4	0.002	4	0.004	4	0.001	4
GDP per Capita (logged)									0.06		0.06		0.07
Government health expenditure								0.577 **	0	0.601 **	1	0.277 **	2
								-	0.00	-	0.00	-	0.00
								0.007 **	3	0.009 **	3	0.003	4
Country level variance	0.097 **	0.01	6	0.095	0.01	8	0.092	5	0.033 **	6	0.034	7	0.047
Country-year level variance							0.002	1					0.003
Individual level variance	0.596 **	0.00	3	0.591	0.00	4	0.591	4	0.596 **	3	0.591	4	0.591
N1 (country-level)	58			55			27		58		55		27
N2 (country-year level)							55						55

Notes: * p<.05, ** p<.01 (two-sided tests); our models are based on 56,712 individual observations; the models in the original study are based on 59,314 observations. The models are named M2 and M3 to fit the naming in the original study.
Source: European Social Survey (2002-2006).

Table 5: Multilevel Regressions with ESS data - Models A-F

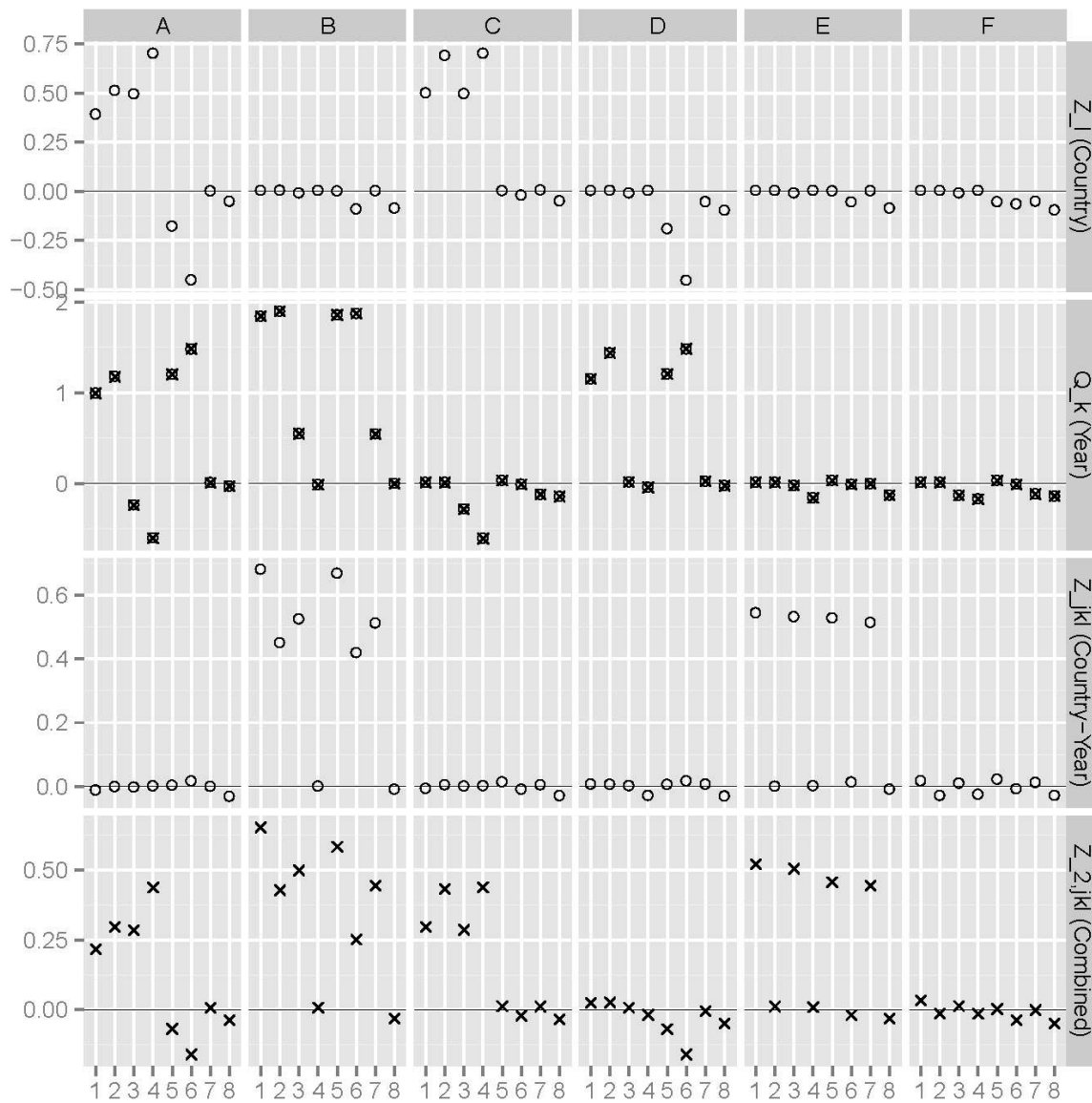
	Model A	Model B	Model C	Model D	Model E	Model F
	b/se	b/se	b/se	b/se	b/se	b/se
<i>Individual-level data</i>						
Left-right	- 0.0853 *** (0.0012)	- 0.0854 *** (0.0012)	- 0.0853 *** (0.0012)	- 0.0853 *** (0.0012)	- 0.0855 *** (0.0012)	- 0.0853 *** (0.0012)
Gender (male=1)	- 0.1225 *** (0.0054)	- 0.1218 *** (0.0054)	- 0.1225 *** (0.0054)	- 0.1225 *** (0.0054)	- 0.1217 *** (0.0054)	- 0.1224 *** (0.0054)
Age	- 0.0031 *** (0.0002)	- 0.0030 *** (0.0002)	- 0.0031 *** (0.0002)	- 0.0031 *** (0.0002)	- 0.003 *** (0.0002)	- 0.0031 *** (0.0002)
Household income	- 0.0798 *** (0.0022)	- 0.0759 *** (0.0022)	- 0.0798 *** (0.0022)	- 0.0795 *** (0.0022)	- 0.0759 *** (0.0022)	- 0.0795 *** (0.0022)
Employment status						
Employed	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
Unemployed	0.0868 *** (0.0126)	0.0886 *** (0.0127)	0.0867 *** (0.0126)	0.0870 *** (0.0126)	0.0876 *** (0.0127)	0.0867 *** (0.0126)
Not in labor force	- 0.0224 *** (0.0064)	- 0.0201 ** (0.0064)	- 0.0224 *** (0.0064)	- 0.0222 *** (0.0064)	- 0.0203 ** (0.0064)	- 0.0223 *** (0.0064)
Education						
ISCED 0-1	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
ISCED 2	0.0643 *** (0.0104)	0.0628 *** (0.0105)	0.0643 *** (0.0104)	0.0651 *** (0.0104)	0.0643 *** (0.0105)	0.0652 *** (0.0104)
ISCED 3	0.0188 (0.0101)	0.0155 (0.0101)	0.0188 (0.0101)	0.0198 (0.0101)	0.0147 (0.0101)	0.0196 (0.0101)
ISCED 4	- 0.0595 ** (0.0216)	- 0.0609 ** (0.0215)	- 0.0595 ** (0.0216)	- 0.0585 ** (0.0216)	- 0.0573 ** (0.0215)	- 0.0579 ** (0.0216)
ISCED 5-6	- 0.1848 *** (0.0106)	- 0.1895 *** (0.0106)	- 0.1849 *** (0.0106)	- 0.1839 *** (0.0106)	- 0.1903 *** (0.0106)	- 0.1842 *** (0.0106)
Bad health	0.0677 *** (0.0063)	0.0704 *** (0.0063)	0.0677 *** (0.0063)	0.068 *** (0.0063)	0.0705 *** (0.0063)	0.068 *** (0.0063)
<i>Country-level data</i>						
GDP/C (in 1,000\$)	- 0.0059 (0.0031)	- 0.0053 *** (0.0007)	- 0.0064 * (0.0031)	- 0.0042 (0.0022)	- 0.0042 ** (0.0015)	- 0.0043 (0.0031)
Gini index	0.0106 (0.0069)	0.0239 *** (0.0022)	0.0104 (0.0069)	0.0257 *** (0.0066)	0.0209 *** (0.0023)	0.0200 ** (0.0064)
Social spending	- 0.0166 * (0.0067)	- 0.0095 *** (0.0022)	- 0.0163 * (0.0067)	- 0.0011 (0.0065)	- 0.0069 * (0.0031)	- 0.0118 (0.0068)
Unemployment rate	- 0.0185 * (0.0075)	- 0.0021 (0.0017)	- 0.0179 * (0.0075)	- 0.0034 (0.0052)	- 0.0002 (0.0017)	- 0.0025 (0.0050)
Constant	4.5484 ***	3.3754 ***	4.5677 ***	3.5723 ***	4.1310 ***	4.2582 ***

	(0.3056)	(0.1118)	(0.3044)	(0.2737)	(0.1410)	(0.2997)
<i>Variance components</i>						
Country		0.1243 0 (0.034 5)		0.0938 0 (0.029 6)	0.0640 3 (0.018 0)	0.0608 8 (0.018 0)
Year			0.0011 3 (0.003 0)		0.0033 4 (0.002 4)	0.0037 2 (0.003 3)
Country-year	0.0703 0 (0.009 8)		0.0692 1 (0.009 9)	0.0073 3 (0.001 4)		0.0069 4 (0.001 3)
Individuals	0.9305 9 (0.003 6)	0.9358 3 (0.003 6)	0.9305 9 (0.003 6)	0.9305 9 (0.003 6)	0.9355 8 (0.003 6)	0.9305 9 (0.003 6)

Notes: * p = .05, ** p = .01, *** p = .001. All models are based on 133,301 individual observations and data from 27 countries and 105 country-years. Models estimated with Stata's xtmixed command.

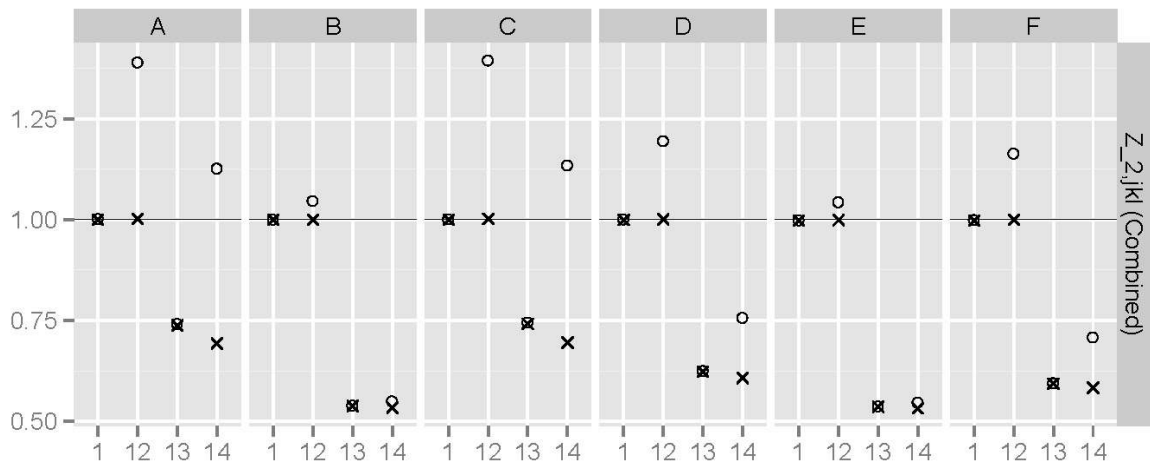
Source: European Social Survey (2002-2010).

Figure 1: Log-optimism of the Estimated Standard Errors of Four Covariates - DGPs 1-8



Notes: See Table 3 for details about DGPs 1-8. O's indicate estimates from models including separate country- and country-year level effects, and X's indicate estimates from models that include only the combined variable. Note then that the estimated coefficients differ: the second approach yields two estimates (for Q_k , a year level covariate, and $Z_{2|k|}$, a country-year level covariate), whereas the first approach estimates fixed effects coefficients for three covariates, including a covariate capturing the between-country effect (Z_1). The correct value of 0 is highlighted with a horizontal line.

Figure 2: Fixed Effects Estimates - DGPs 1, 12-14



Notes: Analogous to DGP 1, the random error variances in DGPs 12-14 are 1 at every level (compare Table 3, DGP 1). O's indicate estimates from models omitting the additional covariate, and X's indicate estimates from models that include the covariate. Note that in DGPs 1 and 13 there is no additional covariate correlated with Z, while there is in DGPs 12 and 14. In DGPs 13 and 14 the between and within effects are 1 and 0.5, respectively, whereas in DGPs 1 and 12 they are both 1. The value of 1 is highlighted with a horizontal line.

Endnotes

1 Multilevel models are also known as random effects, mixed, or hierarchical models. We use the former term as it is the one most often used in the context of survey data analyses.

2 In this paper we deal with RE models. There are, however, alternative ways of accounting for clustered data, which we do not discuss here. Readers who are interested in this are referred to Bell and Jones (2014) for a comparison of random and fixed effects and to King (2014) for a treatment of clustered standard errors.

3 For clarity, when we refer to fixed effects at a given level, we mean covariates measured at that level, not unit dummies intended to capture variance at that level.

4 There are two more logically possible combinations of RE—RE only at the year-level and no RE at all—but these are irrelevant in practice. The latter would not be a multilevel model, and the former would completely ignore the comparative character of the data. In our coding, we found some studies fitting single-level models to multilevel data, in some but not all cases using clustered standard errors; but we do not address those here.

5 We performed the search on the 27th of March, 2014. Detailed information on each stage of our coding process is available upon request.

6 However, 9 out of 14 of these articles did not account for the clustering at the higher levels of geographical clusters at all. They estimated simple panel models but did not include RE for the geographical clusters.

7 The ESS datasets are continuously updated. We have not been able to determine precisely why, but some of the variables which were used in the study appear to have been problematic for some countries, and so have been removed from the available ESS datasets.