Köln Z Soziol (2019) (Suppl 1) 71:99–128
https://doi.org/10.1007/s11577-019-00607-9

Kölner Zeitschrift für Soziologie und Sozialpsychologie

ABHANDLUNGEN

Multilevel Models for the Analysis of Comparative Survey Data: Common Problems and Some Solutions

Alexander W. Schmidt-Catran · Malcolm Fairbrother · Hans-Jürgen Andreß

Published online: 6 May 2019

Abstract This paper provides an overview over the application of mixed models (multilevel models) to comparative survey data where the context units of interest are countries. Such analyses have gained much popularity in the last two decades but they also come with a variety of challenges, some of which are discussed here. A focus lies on the small-*N* problem, influential cases (outliers) and the issue of omitted variables at the country level. Summarizing the methodological literature, the paper provides recommendations for applied researchers when possible or otherwise points to the more detailed literature. Some solutions for the small-*N* problem and omitted variable bias are discussed in detail, recommending the pooling of multiple survey waves to increase statistical power and to allow for the estimation of within-country effects, thereby controlling for unobserved heterogeneity. All issues are illustrated using an empirical example with data from the European Social Survey. The online appendix provides detailed syntax to adopt the presented procedures to researchers' own data.

Online Appendix: http://www.schmidt-catran.de/mixedmodels.html

A. W. Schmidt-Catran (✉)
Institut für Soziologie, Lehrstuhl für Soziologie mit dem Schwerpunkt Methoden der quantitativen empirischen Sozialforschung, Goethe-Universität Frankfurt
Theodor-W.-Adorno-Platz 6, Campus Westend, 60323 Frankfurt am Main, Germany
E-Mail: alex@alexanderwschmidt.de

M. Fairbrother
Department of Sociology, Umeå University
Norra Beteendevetarhuset, Umeå universitet, 901 87 Umeå, Sweden
E-Mail: malcolm.fairbrother@umu.se

H.-J. Andreß
Fakultät für Wirtschafts- und Sozialwissenschaften, Institut für Soziologie und Sozialpsychologie, Lehrstuhl für empirische Sozial- und Wirtschaftsforschung, Universität zu Köln
Albertus-Magnus-Platz, 50923 Cologne, Germany
E-Mail: hja@wiso.uni-koeln.de

Keywords Mixed models · Multilevel models · Small-*N* problem · Influential cases · Omitted variable bias

Mehrebenenmodelle zur Analyse von vergleichenden Umfragedaten: Häufige Probleme und ausgewählte Lösungsansätze

Zusammenfassung Die vorliegende Arbeit bietet einen Überblick über die Anwendung von Mehrebenenmodellen auf international vergleichende Umfragedaten. Mehrebenenanalysen, in denen die relevanten Kontexteinheiten Länder sind, haben in den letzten 2 Jahrzehnten eine weite Verbreitung gefunden, sind allerdings aus statistischer Perspektive in einigen Aspekten problematisch. Dieser Artikel zielt auf einige der Probleme ab, die bei der Anwendung von Mehrebenenanalysen auf internationale Umfragedaten auftreten. Ein Fokus liegt dabei auf dem small-*N*-Problem, einflussreichen Fällen („Ausreißern") und dem Problem unbeobachteter Heterogenität auf der Länderebene. Dieser Beitrag bietet eine Zusammenfassung der methodischen Literatur zu Mehrebenenmodellen und versucht, in Forschung Tätigen möglichst konkrete Empfehlungen zu geben oder – wo dies nicht möglich ist – auf die tiefergehende Literatur zu verweisen. Lösungsansätze für das small-*N*-Problem und das Problem unbeobachteter Heterogenität werden im Detail diskutiert. Aus dieser Diskussion ergibt sich die Empfehlung, vorhandene Wellen international vergleichender Umfragedaten zu poolen. Zur Illustration verwendet dieser Artikel ein empirisches Beispiel auf Basis der Daten des European Social Survey. Der Online-Anhang enthält zu diesen Beispielen eine detaillierte Syntax, die sich leicht für andere Daten und Forschungsfragen anpassen lässt.

Schlüsselwörter Gemischtes Modell · Mehrebenenmodelle · Small-*N*-Problem · Ausreißer · Unbeobachtete Heterogenität

1 Introduction

Multilevel models, also known as random effects, hierarchical, or mixed models, are regression models for the analysis of hierarchical data. Such models can be applied to a wide variety of data structures, but applications to two types of data are particularly common in the social sciences: (1) panel data, where measurement occasions are nested in persons or some other unit of analysis (e. g. firms, nations); and (2) datasets where the primary units of analysis (e. g. survey respondents, employees, students) are nested in higher-level social groups (e. g. nations, companies, schools). This paper focuses on the latter type, and particularly on the decisions confronting researchers analyzing comparative survey data, though it also considers insights developed in the tradition of panel data analysis.

Due to the vast increase in the availability of comparative surveys during the last two decades, the expansion of computational power, and improvements to statistical software, multilevel models have become a commonly used tool of social science. To illustrate the point, Fig. 1 shows the share of multilevel analyses out of all articles appearing in the *European Sociological Review* (ESR) from 2000 to 2016;

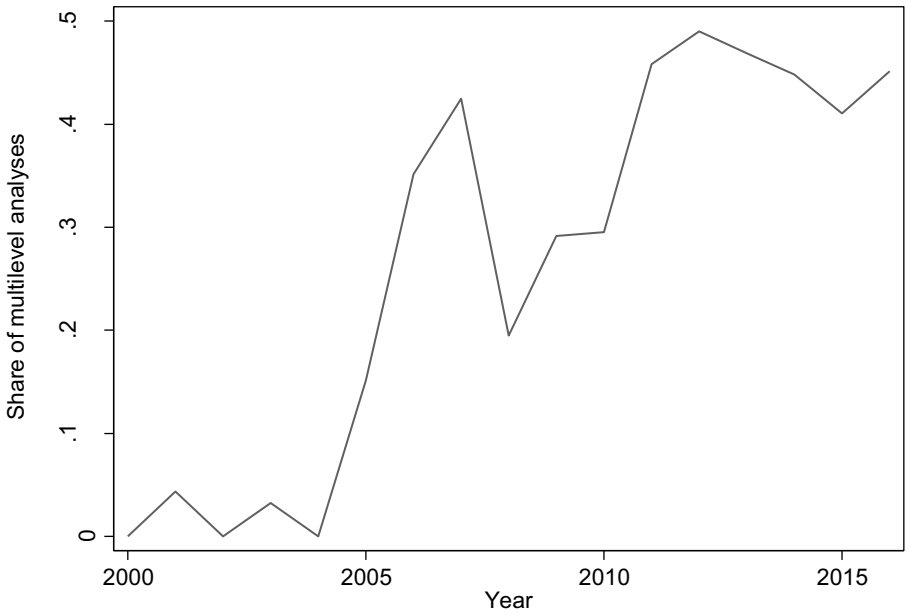

Fig. 1 Share of multilevel analyses from all publications in *European Sociological Review* (ESR). Notes: Based on a keyword search for the term "multilevel" in the search engine of ESR (on October 10, 2017) and the total number of articles published between 2000 and 2016

the proportion has reached almost 50%. Specifically with respect to comparative survey datasets (i.e., surveys conducted in multiple countries simultaneously, such as the European Social Survey or World Values Surveys), multilevel models are a popular analytical tool because they help identify how individual outcomes like attitudes and behaviors vary according to social context. All the social sciences take an interest in how people's economic, social, political, or institutional circumstances shape their lives.

In the face of the dramatically expanding popularity of multilevel modelling, and the creative application of such models to new kinds of data and research questions, methodologists have started to point out problems and challenges in specific analyses and common research practices (e.g., Bryan and Jenkins 2016; Heisig and Schaeffer 2018; Schmidt-Catran and Fairbrother 2016; Te Grotenhuis et al. 2015). Drawing on this literature, this paper discusses some issues particularly relevant for analyses of comparative survey data: statistical inference with nonrandom samples; the problem of having only a small number of higher-level units; and issues of omitted variable bias. These issues are not unique to analyses using multilevel models, but are rather general problems for all kinds of regression techniques, and therefore where appropriate we bring in insights from more general literature.

Throughout the discussion, to provide a concrete illustration of the general points we make, the paper uses a running example inspired by a recent study by Te Grotenhuis et al. (2015). Investigating the relationship between social security and religious involvement, Te Grotenhuis et al. (2015) demonstrate, in their words, "the danger of

testing hypotheses cross-nationally." Substantively, their study tests whether state-provided social security, along with general increases in economic wealth, can substitute for some of the benefits to individuals that come from religion. For a detailed theoretical treatment of this hypothesis, we refer readers to the paper by Te Grotenhuis et al. (2015) and the literature cited therein. Methodologically, Te Grotenhuis et al. (2015) used Eurobarometer data, but we employ data from the European Social Survey (ESS; 2016), like a prior study on the same subject by Immerzeel and Tubergen (2013). All analyses in this paper can be replicated using the Stata data set and do-file provided in the online appendix.[1]

We will focus on linear multilevel models for continuous dependent variables. We begin with a very brief introduction to these models and their assumptions. For ease of presentation, we will from now on always refer to the example of individuals (at level 1) nested in countries (at level 2).

1.1 A Very Brief Introduction into Multilevel Models

A multilevel model for continuous dependent variables is a generalization of the linear regression model, which includes a separate error component at each of its levels and may be written as

$$y_{ji} = \beta_0 + \beta_1 x_{1ji} + ... + \beta_k x_{kji} + \gamma_1 z_{1j} + ... + \gamma_l z_{lj} + u_j + e_{ji},$$

where the index i indicates individuals and j indexes countries. From left to right, y_{ji} is an individual-level outcome (e. g. church attendance), and the model includes 1 to k individual-level variables x (e. g. age, education), with corresponding coefficients β, and 1 to l country-level variables z (e. g. social spending, GDP/capita where GDP is gross domestic product), with the coefficients γ. These coefficients are conventionally also referred to as fixed effects. In addition, the model also includes random effects (or error terms) at the individual (e_{ji}) and the country level (u_j), both of which are assumed to be normally distributed with a mean of zero and a constant variance and to be uncorrelated with each other and with the observed variables. Where the purpose of the analysis is to identify a causal relationship, the latter assumption is called the *exogeneity* assumption and is crucial for the estimation of unbiased fixed effects. The variances of the error terms are estimated, with the term u_j capturing the country-level disturbances from the overall intercept β_0. Each individual element of u_j is called a random intercept.

In fitting multilevel models, it is common for researchers to calculate the intraclass correlation (ICC): the share of the total unexplained variance attributable to the higher level. The formula for this is $\rho = \sigma_u^2 / \left(\sigma_u^2 + \sigma_e^2\right)$, where σ_u^2 and σ_e^2 are the variances of the individual- and country-level random effects, respectively (Hox 2010, p. 15). In an empty model—a model that includes no observed independent variables—the ICC indicates what proportion of the overall variance is at the country level, a figure equivalent to the average correlation of observations within countries. If it were zero, the observations would not violate the assumption of independence,

[1] The online appendix is available at www.schmidt-catran.de/mixedmodels.html.

there would be no intercountry differences to explain, and a multilevel model would not be necessary.

Considering our research example, we can examine the degree to which religious involvement varies across countries, ahead of explaining that variation by social security and other variables. We follow Grotenhuis et al. (2015) in operationalizing religious involvement as church attendance, and in their treatment of this variable as interval-scaled (such that a linear model can be estimated). Using the ESS wave from 2014, we find $\rho = 0.335/(0.335 + 2.030) = 0.142$. Thus, 14.2% of the total variance in church attendance is attributable to the country level (Table 4 in the appendix describes the sample used for this analysis).

The model above can be extended and made more flexible, allowing not only for the intercept β_0 to vary cross-nationally, but also for any individual-level variable's effect to vary between countries. Such a model is often called a random intercept and random slope model:

$$\begin{aligned} y_{ji} =& \beta_0 + \beta_1 x_{1ji} + ... + \beta_k x_{kji} + \gamma_1 z_{1j} + ... + \gamma_l z_{lj} + \\ & u_{0j} + u_{1j} x_{1ji} + ... + u_{kj} x_{kji} + e_{ji} \end{aligned}$$

The random effects u_{1j} to u_{kj} are country-level variances that capture the deviation of country-specific slopes from the average effects across all countries (β_1 to β_k).[2] Thereby the model explicitly allows for heteroscedasticity due to effect heterogeneity in individual-level variables. The random effects at the country level—random intercepts and slopes—are assumed to have a multivariate normal distribution and be independent of the idiosyncratic error term e_{ji}.

The covariances between random intercept and slopes, however, are not or rather should not be assumed to be zero (Hox 2010, p. 13). This means we generally estimate a variance-covariance matrix for the random effects (intercepts and slopes) of the form

$$\Sigma_u = \begin{pmatrix} \sigma^2_{u0} & \cdots & \sigma_{u0}\sigma_{uk} \\ \vdots & \ddots & \vdots \\ \sigma_{u0}\sigma_{uk} & \cdots & \sigma^2_{uk} \end{pmatrix},$$

where the diagonals of this matrix describe the variances of random effects and the off-diagonals include the covariances between each pair of random effects. The number of unique entries in this symmetric matrix, together with the number of country-level variables in the fixed part of the model, constitutes the total number of parameters estimated from country-level information. For example, a model including two country-level variables (e.g. social spending and GDP/capita) and three random slopes of individual-level variables (e.g. gender, age, education) will estimate 12 country-level parameters in total: two country-level fixed effects, four random effect variances (intercept plus three slopes) and six covariances between

[2] Note that the country-level random effects now have an additional subscript (0,1, ... ,*k*), indicating to which fixed effect the random effect belongs.

Table 1 Random intercept models of church attendance, European Social Survey (ESS) 2014

Variable	M0		M1		M2	
Individual-level variables						
Urban vs. rural	–	–	0.0669	***	0.0669	***
Education (in years)	–	–	–0.0156	***	–0.0156	***
Subjective income	–	–	–0.0120	–	–0.0123	–
Male (ref = female)	–	–	–0.2077	***	–0.2076	***
Age	–	–	0.0091	***	0.0091	***
Country-level variables						
Social spending (% of GDP)	–	–	–	–	–0.0465	–
GDP/capita	–	–	–	–	0.0000	–
Average urban vs. rural	–	–	–	–	0.4394	–
Average education	–	–	–	–	–0.0996	–
Constant	1.4620	***	1.1409	***	2.7045	–
Variance components						
Country level	0.3348	***	0.3265	***	0.3005	**
Individual level	2.0231	***	1.9662	***	1.9662	***
Statistics						
N (Country)	20		20		20	
n (Individual)	37,028		37,028		37,028	

See text for explanation of M0–M2
*$p < 0.05$, **$p < 0.01$, ***$p < 0.01$ (two-sided tests). All models are estimated via Restricted Maximum Likelihood. Models based on ESS data 2014 (compare Table 4)
GDP gross domestic product

them.[3] We will return to this point when discussing the small-*N* problem; suffice to say here that it can be hard not to ask too much of the data while still accounting for an adequate number of fixed and random country-level effects.

Table 1 presents a first analysis of the example data, using a single wave of the ESS.[4] It shows the basic stepwise procedure usually applied with multilevel models. Model M0 is an empty model which is used to decompose the total variance into its individual- and country-level components. As already noted, 14.2% of the variance is at the country level. The next step, as is typical, adds the individual-level variables to the model (M1). Older people, people living in rural areas, women, and people with less education attend religious services more often. Subjective income does not have a significant effect.

By adding individual-level variables first, the analysis reveals how much of the country-level variance can be explained by individual-level differences: $1 - (0.3265/0.3348) \approx 0.025$. This is, 2.5% of the differences between countries can be explained by differences in the populations of the individuals living in those coun-

[3] A symmetric variance-covariance matrix of size *m* contains $m \cdot ((m+1)/2)$ unique entries, *m* of which are variances and the rest being covariances.

[4] A detailed description of all involved variables, their descriptive statistics and correlations, can be found in the online appendix to this paper.

tries. This is often called a *compositional effect* and in this application only a small fraction of the between-country variance can be explained by differences in composition, which means there is substantial variance left that is due to country-level effects. If most of the variance between countries could be explained by compositional effects, we would have to conclude that any differences between countries are not related to contextual effects—only to characteristics of the individuals making up the populations of these countries.

The third step (M2) adds country-level effects, which after controlling for compositional effects can be interpreted as contextual effects. These reduce the unexplained country-level variance from Model M1 by about 8% ($1-(0.3005/0.3265) \approx 0.080$). Social spending (as % of GDP) has the hypothesized negative effect on church attendance, consistent with the results of Immerzeel and Tubergen (2013). However, in contrast to their analysis, the effect of social spending is not significant, which may not be a surprise given that we use 20 observations to estimate five parameters (four fixed and a random effect).

A fourth step could be to test for random slopes and a fifth one the inclusion of cross-level interaction effects, which might explain the variation in individual-level effects identified in step four. (For a detailed description of the stepwise procedure see Hox 2010, p. 54 ff.). Following Te Grotenhuis et al. (2015) and Immerzeel and Tubergen (2013) we are not interested in cross-level interactions and therefore stop here.[5]

This has clearly been a very brief introduction, but it should have served the purposes of introducing some notation and core ideas, and starting some analysis of the example dataset. For a detailed introduction to multilevel models, readers may wish to consult one of the classic introductory textbooks by Hox (2010) or Snijders and Bosker (2012). Rabe-Hesketh and Skrondal (2012) provide an easily accessible introduction into multilevel models using Stata. Gelman and Hill (2007) discuss multilevel models in both frequentist and Bayesian frameworks, using the software packages R and BUGS.

2 Challenges in Analyses of Comparative Survey Data

Multilevel analyses of comparative survey data are not without their complications. Measurement equivalence with respect to latent variables, for example, can be a limitation—as explained in the paper by Cieciuch et al. in this special issue. Setting aside problems of measurement, however, here we address a different set of issues.

First, the countries included in international surveys are never random samples, but are instead selected or self-selected in ways that make them, effectively, convenience samples (Ebbinghaus 2005). This raises questions about the justifiability of statistical inferences to a larger population of countries, and about the use of infer-

[5] But see Heisig et al. (2017) who argue for the inclusion of random slopes even if the research interest is not in cross-level interactions, i.e. in explaining differences in individual-level effects by country-level characteristics. Barr et al. (2013) and Bell et al. (2019) also demonstrate and discuss the importance of random slopes.

ential statistics generally (see Goerres et al. 2019).. Second, the number of countries included in such surveys is typically rather small. Most international surveys include about 30 countries (e. g. European Social Survey [ESS]; European Union Statistics on Income and Living Conditions), and only a few include more than about 50 (such as by combining samples from the World Values Surveys [WVS] and European Values Studies [EVS]). Many studies analyze an even more limited number of countries because right-hand-side national-level variables are often unavailable for some countries (Bryan and Jenkins 2016, p. 3). This increases both the selectivity of the sample (Ebbinghaus, 2005: p. 136) and the severity of the small-*N* problem. Third, in a model aiming at identifying a causal relationship, the small degrees of freedom at the country level limits the number of higher-level control variables that can be included (see Goldthorpe 1997, p. 5 f.; Jaeger 2013). We discuss each of these issues in turn.

2.1 Nonrandom Country-level Sampling in International Surveys

From the point of view of some researchers, inferential statistics are only applicable to random samples, which leaves rather unclear the statistical status of analyses conducted on, in effect, convenience samples of countries. Some researchers conclude that inferential statistics are completely meaningless in these settings; others argue that the use of inferential statistics is justified even with these nonrandom samples (compare Ebbinghaus 2005 and Babones 2013, 107 ff.).

When observations on entire countries are the units of analysis, as in the analysis of pooled time-series cross-section data, the research community tends not to object to the use of inferential statistics. That is true even though the nonrandom sampling of countries prohibits the straightforward generalization of findings to a larger population of countries; instead "all inferences of interest are conditional on the observed units" (Beck 2001, p. 273).

While samples of countries in international surveys are clearly not random—and therefore country-level effects must be viewed as conditional on the specific sample of countries—at the very least individuals within countries are sampled at random.[6] Therefore individual-level results should be generalizable within countries. However, individual-level effects in multilevel models are not only identified by variation within countries, but also by between-country variation (see Bell et al. 2018; Andress et al. 2013, particularly p. 157 ff.). This also implies that inference to the populations within countries may be problematic. One way of addressing this problem is to group-mean center the individual-level variables, stripping them of any country-level variation (Hox 2010, p. 68 ff.; Bell et al. 2018; see Fairbrother 2016 for an applied example).[7] Enders and Tofighi (2007) suggest doing this if the interest is purely in individual-level relationships, though multilevel models are typically employed because of a specific interest in country-level effects or their interactions with individual-level variables. However, if the interest is really just in individual-

[6] The issue of nonrandom missing values, i. e. sample selection effects at the individual level, is left aside here.

[7] This is equivalent to the introduction of country-dummies, i. e. country fixed effects.

level effects, other modelling techniques may be better suited (Bryan and Jenkins 2016).

There is an informal working consensus in the literature that inferential statistics are also relevant at the country level, despite the fact that the countries included in international surveys are not selected at random from the population of all countries. The basic argument for this is that there are several other relevant sources of random variation, aside from sampling errors (e. g. measurement errors, omitted variables), which justify the usefulness of *p*-values for separating real effects from random noise.

What does this imply for the research example? The ESS data used here are obviously not a random sample of countries and certainly cannot be used to generalize results to the world population of countries in a statistical sense (see Table 4 for a sample description). The original data set from the ESS included 32 countries[8] and covered most EU member countries. So, one might think that models based on this data should allow to make statements about EU member countries. Due to missing data for social spending and/or GDP per capita, however, some countries were excluded from the analysis. If the missing observations were truly random, the data would allow for generalization to the population of EU member countries.[9] However, the excluded countries are Bulgaria, Cyprus, Croatia, Lithuania and the Ukraine, seemingly not a random set of countries.

2.2 The Small-*N* Problem

We coded articles with multilevel analyses in the *European Sociological Review* and found 103 such analyses using countries as contextual units. In those analyses, the average number of countries is 22.6 (Min = 9, Max = 78). Setting aside the issue of nonrandom sampling, then, what are the implications of using such small country-level samples in multilevel models of comparative survey data?

First, with higher-level *N*s in this range, the estimated coefficients of country-level variables will often be quite sensitive to single (outlying) countries (Wilkes et al. 2007; Van der Meer et al. 2010). Figure 2 tests this possibility for the example data. It presents the simple bivariate relationship between church attendance and social spending (as % of GDP) using the complete ESS data (rounds 1 to 7, compare Table 4), aggregating each variable to the country level. The set of grey lines describes the bivariate relationships when each country is excluded from the sample one at a time; the black line indicates the relationship in the full data. In terms of correlations the strength of the relationship in the full sample is –0.34. When leaving out each country once, it varies between –0.27 (leaving out Turkey) and –0.41 (leaving out Estonia), a substantive difference of about 52%.

One can take two perspectives on this. On the one hand, we can accept that any statistical inference is conditional on the sample and thus it is to be expected that different samples will provide results that deviate from each other by more than

[8] The data set has been obtained from the cumulative data wizard, which does exclude Albania, Kosovo and Latvia.

[9] Ignoring for now the fact that two EU members are not in the sample: Malta and Latvia.

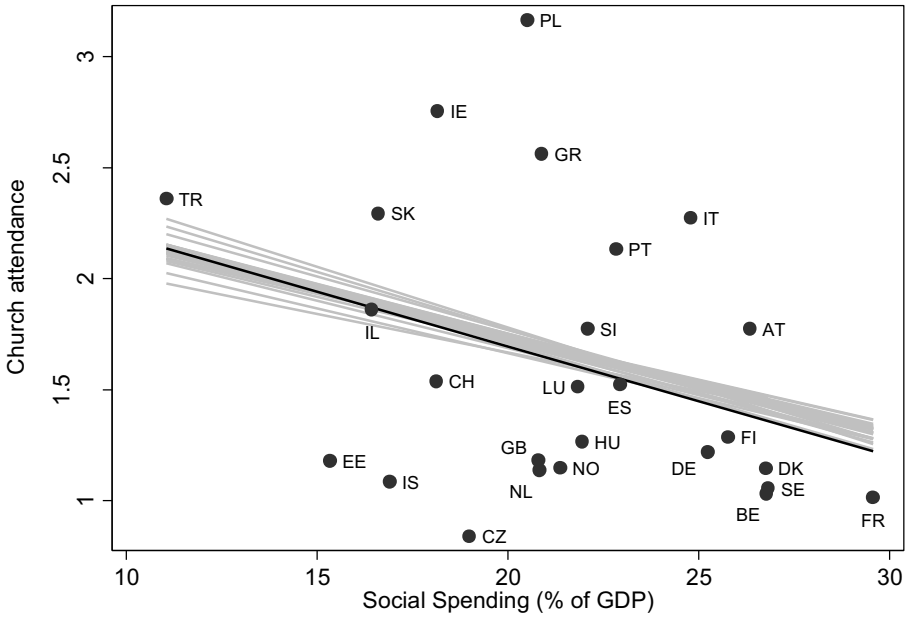

Fig. 2 Bivariate country-level relationships between social spending and church attendance. Notes: Based on ESS data 2002–2014 (compare Table 4). The *black line* represents the association in the full sample, while the *grey lines* represent the associations when leaving out each country one at a time. *AT* Austria, *BE* Belgium, *CH* Switzerland, *CZ* Czech Republic, *DE* Germany, *DK* Denmark, *EE* Estonia, *ES* Spain, *FI* Finland, *FR* France, *GB* Great Britain, *GR* Greece, *HU* Hungary, *IE* Ireland, *IL* Israel, *IS* Iceland, *IT* Italy, *LU* Luxembourg, *NL* Netherlands, *NO* Norway, *PL* Poland, *PT* Portugal, *SE* Sweden, *SI* Slovenia, *SK* Slovakia, *TR* Turkey

what could be explained by sampling error. On the other hand, the model parameters ought to describe the data in the best possible way. In some cases, outliers can have such strong influences that the regression line primarily describes the position of the outlier relative to the rest of the countries, rather than the relationship in the bulk of the data. Van der Meer et al. (2010) provide such an example where a strong positive relationship between church attendance and volunteering completely dissolves once outliers are considered (also see Hox 2010, p. 29).

Investigating outlying cases can be done graphically by means of scatter plots, as in Fig. 2. But scatter plots show only simple bivariate relationships of aggregated data and it may be hard to decide which countries are too influential.[10] An alternative are outlier statistics such as Cook's Distance (Cook 1977) or DFBETAs (Belsley et al. 1980, p. 13), which can also be applied to multilevel models (Snijders and Berkhof 2008, p. 157). Later, in Sect. 4, we demonstrate how to apply these outlier statistics to multilevel models. For now, we simply note that the bivariate cross-sectional relationship between aggregated church attendance and social spending is in line with our expectations: higher spending is associated with less religious

[10] See Bowers and Drake (2005) for more information on how to use exploratory data analysis and visualization when the number of level 2 units is small.

involvement. While the estimated relationship is dependent on the specific countries in the sample, ranging from –0.27 to –0.41, this influence may not be regarded as overly problematic since the complete range of values confirms our theory.

Second, while all available estimation techniques for multilevel models (e. g., Full Maximum Likelihood [FML], Restricted Maximum Likelihood [RML]) are consistent, meaning that they converge to the true parameters with increasing sample size, their behavior in small samples is sometimes problematic (Hox 2010, p. 40 ff.). This issue has motivated several methodological studies asking variations on "how many countries do you need for multilevel modelling?" (Stegmueller 2013; also see Maas and Hox 2005; Bell et al. 2014; Bryan and Jenkins 2016; Heisig et al. 2017; Elff et al. 2016). Such studies have also examined how different estimators behave under conditions of varying sample sizes, violations of the normality assumptions, and other data characteristics.

Both FML and RML, the most commonly applied estimators (Hox 2010, p. 40), provide unbiased point estimates of the fixed effects in *linear mixed models* but the variance components and their standard errors (SEs) are underestimated in small samples. Due to the uncertainty in the random part of the model, the SEs of the fixed effects are also biased downwards, resulting in unclear distributions of test statistics and the risk of performing anticonservative tests[11] (Bryan and Jenkins 2016, p. 7; also see Elff et al. 2016, p. 14 ff. for some solutions). The same biases are found with nonlinear multilevel models with the additional caveat that the unbiasedness of fixed effects coefficients cannot be clearly demonstrated for these models (Bryan and Jenkins 2016, p. 7 f.).

The small-sample bias appears to be much stronger with FML than with RML (Hox 2010, p. 41). In fact, RML was introduced to deal with the FML bias in variance component estimation (Patterson and Thompson 1971). Nevertheless, Maas and Hox (2005) find somewhat substantial biases of RML with small samples. Most studies, however, find very small or nonexisting biases with RML even if the country-level N is as small as 10 or 5 (Bryan and Jenkins 2016; Browne and Draper 2000; Elff et al. 2016). With FML, in contrast, the bias can be quite substantial with small samples at the country level (Elff et al. 2016; Browne and Draper 2000).

Should one always prefer RML over FML then? FML has one clear advantage vis-a-vis RML, which is that it allows the use of likelihood-ratio tests (LR tests) to compare nested models (Hox 2010, p. 41).[12] Such comparisons can be very useful in the process of model building and may also be helpful for testing hypotheses. Thus, there is a trade-off between RML and FML: If the bias of FML estimates is negligible, FML may be preferred over RML. Above an ICC of 0.142 was obtained from the example data on church attendance. This model was estimated with RML. Using FML, the ICC is estimated to be 0.136. As expected the FML estimates yield a smaller variance at the country level but the difference may be regarded as trivial.

[11] With anticonservative tests, the risk of falsely rejecting the null hypothesis of no effect increases. In other words, results look too significant.

[12] To be precise, RML does also allow to compare nested models but only if they differ in their random but not in the fixed part.

This is consistent with a recent simulation study by Elff et al. (2016, p. 13 ff.), who show that the bias of FML compared to RML is substantial with fewer than 15 countries but relatively unimportant with 20 or more. Nevertheless, we suggest that instead of applying simple rules of thumb, researchers should compare the results of both methods to decide whether the bias of FML can be ignored. Formulating a rule of thumb is difficult because the performance of any estimator is highly dependent on the specifics of the data and the complexity of the model fitted to them (Bryan and Jenkins 2016, p. 8).

In addition to FML and RML, there are several other estimators for multilevel models available: Generalized Least Squares (GLS), Generalized Estimation Equation (GEE), and Bayesian methods. GLS is asymptotically equivalent to FML but in practice often less efficient (Hox 2010, p. 42 f.). GEE and cluster robust SEs can be a remedy against too optimistic (underestimated) SEs but also involve the risk of obtaining overestimated SEs (Hox 2010, p. 262 f.), which are to be avoided given that the statistical power to estimate country-level effects is rather small anyway. With violated distributional assumptions, which can be a consequence of a small *N* at the country level, bootstrapping can reduce the bias in SEs but it is implemented only in a few statistical software packages, is computationally quite demanding, and is not *per se* useful with small samples (Hox 2010, p. 264 ff.). For now valid bootstrapping with multilevel models is implemented only in MLwiN.

Finally, there is the option to turn away from classical frequentists statistics and use Bayesian methods. Obviously, this paper does not offer the space to deal with Bayesian methods in any detail. Readers who are interested in Bayesian multilevel modelling may want to start with Jackman (2009), who gives a general introduction into Bayesian modelling and treats multilevel models in Chap. 7. Hox (2010) has a large section on Bayesian multilevel modelling (p. 271 ff.); Gelman and Hill (2007) and Draper (2008) may also be good starting points.

In a nutshell, frequentists view the population *parameter as* an unknown but *fixed* quantity, which they estimate from data. The uncertainty in the estimate results from the sampling distribution, i. e. the distribution of the parameter in an indefinite number of samples. Bayesians view the *data as fixed* and the parameter of interest as an unknown quantity that must be described by probabilistic statements and can always be updated by data. This leads Bayesians to formulate a prior distribution, which reflects the belief, or rather (un)certainty, about the parameter before seeing the data. The data then is used to update the prior distribution by conditioning it on the observed data, resulting in the so-called posterior distribution. This posterior distribution, the result of the analysis, characterizes the researcher's new beliefs about the parameter, in light of the prior distribution and the likelihood of the data.

With large *N*s and uninformative priors—priors that do not favor any specific parameter region—Bayesian estimates are identical to ML estimates. There is some controversy about the question of whether a Bayesian approach deals better with the small-*N* problem than frequentist analysis does. Stegmueller (2013) claims that Bayesian methods have an inherent advantage over frequentists methods when it comes to the analysis of hierarchical data with few clusters. Elff et al. (2016) disagree. In our reading of the literature, the unbiasedness of Bayesian methods with small *N*s is more straightforward than it is for the frequentist approach, within which

special adjustments and estimation methods are needed for small samples (compare Elff et al. 2016). On the other hand, some literature suggests that seemingly uninformative priors can result in biased Bayesian estimates when the sample size is small (Gelman 2006; Van Erp et al. 2017). In sum, there does not seem to be a general advantage of Bayesian methods over frequentist approaches.

It is a different game, of course, if a researcher has useful prior information on parameters, in which case the Bayesian approach can be recommended. But we have yet to see a convincing implementation of a model using informative priors in the context of comparative survey data. It is telling that out of the (just) six Bayesian multilevel analyses published in ESR since 2000[13] none used (true) informative priors—one analyses (Sutton 2012) implemented so-called skeptical priors, which drag coefficients slightly towards zero to create conservative tests.

2.3 Omitted Variable Bias

To identify a causal effect of a variable x on y, any alternative explanation for an association between them must be ruled out. In experiments this is of course achieved by randomization. With observational data, it must be done by partialing out the effects of any variable that is a cause of both y and x. Technically, the omission of a variable which affects y and is related to x violates the exogeneity assumption and therefore results in biased coefficient estimates (Wooldridge 2013, p. 88 ff., also see 45 ff.). This very basic insight is no different for multilevel models (Kim and Frees 2006).

However, with multilevel models fitted to comparative survey data, the small-N problem makes the issue of omitted variables even more delicate: First, as we argued above, the limited degrees of freedom at the country level create a trade-off between the need to control for all necessary variables and respecting the limits of what the data can do (Heisig et al. 2017). Second, country-level characteristics of interest are often strongly correlated with each other and with necessary control variables (Babones 2013, p. 94 ff.).[14] Additionally, any attempts to control for an adequate number of country-level (fixed and random) effects are practically limited far below the theoretically absolute limit set by the country-level degrees of freedom because multilevel models tend to run into convergence problems if they include too many covariates at the country level (Heisig et al. 2017, p. 823 f). This combination of high multicollinearity coupled with few degrees of freedom will often result in inefficient estimates and thereby create the temptation to ignore important variables (Arceneaux and Huber 2007).

This has led to a questionable practice in applied research where many researchers make arguments like this: “If all country-level variables are included at the same

[13] Brännström (2008); Sutton (2012); Stadelmann-Steffen (2012); Stegmueller et al. (2012); Giger (2012); Mewes (2014).

[14] In the example data, the country-level variables are not too strongly related. The average (absolute) correlation across the four variables amounts to 0.31 (min = 0.19, max = 0.47), so collinearity is not a pressing issue. However, it is much stronger than the average (absolute) correlation across the individual-level variables which is 0.09 (min = 0.01, max = 0.25).

time, nothing is significant; so, I test and/or control each variable separately".[15] From a causal identification standpoint this strategy is problematic. This is not to say that researchers should include any (control) variable they can think of. In contrast, the model building strategies developed in the framework of directed acyclical graphs provide very good guidance on which variables need to be included in a model and which *not* (for an overview, see Elwert 2013). But to control only piecewise—one variable at a time— is certainly not a good strategy to identify causal effects.

Third, with countries it is arguably very difficult to operationalize all relevant factors (Babones 2013, Chap. 3). Thus, biased estimates due to omitted variables are quite likely outcomes in the analysis of comparative survey data—maybe even more so than with plain individual-level analyses, where the available degrees of freedom tend to be much higher, and measurement in many domains, specifically of latent variables, is arguably easier (Fontaine 2015). There are good reasons to be cautious before concluding that the model has no omitted variables, even if we can include all *available* variables without running into issues of nonconvergence or multicollinearity. After about a decade of elated investigations into country effects, social science researchers started to increasingly worry about such unobserved heterogeneity (for examples, see Fairbrother 2013, p. 911; Jaeger 2013, p. 156; Wulfgramm 2014, p. 263; Schmidt-Catran 2016, p. 124; Te Grotenhuis et al. 2015, p. 644; Finseraas 2012, p. 167).

3 Some Solutions and Caveats

With just a few countries in cross-sectional analyses, and few degrees of freedom at the country level, models may yield imprecise estimates of country-level effects. One way to get more variation at the country level, however, is to observe the same countries multiple times. And many international surveys have now been fielded on multiple occasions (e. g. ESS, ISSP, EVS, WVS), providing an opportunity to pool comparative survey data across time. The resulting data structure may be called comparative *longitudinal* survey data (Fairbrother 2014) and promises to not only increase statistical power but also to provide less biased estimates in the presence of unobserved country-heterogeneity. The former is a direct result of pooling across time, while the latter can be achieved by the identification of country-level effects via within-country variation, i. e. changes of country-level variables over time.

3.1 Comparative Longitudinal Survey Data

As Schmidt-Catran and Fairbrother (2016, p. 26) show in their literature review, many researchers have attempted to apply multilevel models to comparative longitudinal survey data. But they also demonstrate that there are right and wrong ways of analyzing such data, and previous studies have often used problematic specifications.

[15] An example of such a paper is Semyonov et al. (2006, p. 437): "Because of restrictions associated with the limited degrees of freedom at the country level, only three hierarchical linear model equations are estimated [...], with each equation including only one country-level variable."

Specifically, the introduction of a longitudinal dimension into the data creates an additional level in the hierarchical structure of the data, and this level must be accounted for to obtain unbiased SEs. In other words, incorrectly specifying the statistical model can lead to significance levels that are not actually supported by the data. Moreover, Schmidt-Catran and Fairbrother (2016, p. 30, 34) also demonstrate that a failure to model the correct random effects structure may not only yield overly optimistic SEs, but also biased coefficient estimates.

So, what is the correct hierarchical structure for a given analysis? This depends on two questions: First, at which levels are the variables measured and, second, at which levels is there variation in the data? Comparative longitudinal survey data can be viewed as having four levels: countries, survey waves (typically years, which will be used synonymously from here), combinations of countries and waves (here called country-years), and individuals. Thus, there are potentially three levels above the individuals (years, countries and country-years). At each of these levels there may be variation, meaning the observations within these clusters can be dependent. For example, individuals within the same countries are more similar than individuals from different countries; but they may even be more similar if they are observed in the same year. Alternatively, individuals observed in the same year may be more similar than individuals observed in different years, even if they are observed in different countries. Such variation needs to be accounted for by random or fixed effects. The latter can be done via the introduction of dummy variables for the clusters.

Including such dummies, however, takes up all the degrees of freedom at that level, which means no variables can be included at this level.[16] Thus, for each variable of interest, there needs to be a corresponding level in the random part of the model. This leaves only levels as candidates for cluster-dummies at which no variables of interest are measured. The final question then is the following: At what level is a variable measured? In the simple two-level model from above, with cross-sectional data, this question is easy to answer. Individual-level variables (e. g. age, gender) are measured at the individual level and country characteristics (e. g. social spending, GDP/capita) are measured at the country level.

When a longitudinal dimension comes into play, this question becomes more complicated. By definition, a cluster-level variable must be constant within clusters. Thus, a country-level variable that changes over time, like social spending, is not a country-level variable. For this reason, Schmidt-Catran and Fairbrother (2016) argue that comparative longitudinal survey data are—in most cases—best analyzed with the following model:

$$y_{jti} = \beta_0 + \sum_{t=1}^{T} \delta_t D_t + \beta_1 x_{1jti} + ... + \beta_k x_{kjti} + \gamma_1 z_{1jt} + ... + \gamma_l z_{ljt} + u_j + u_{jt} + e_{jti}$$

[16] Technically, there is perfect collinearity between country-level variables and country-dummies.

This is a hierarchical three-level model with individuals (i) nested in country-years (jt) nested in countries (j). The term u_j captures (unexplained) variance between countries and u_{jt} accounts for the (unexplained) variance within countries over time. The potential variance at the year level is not modelled via random effects but with year-dummies ($\sum_{t=1}^{T} \delta_t D_t$). This model allows for the inclusion of time-constant country-level variables (e. g. legal tradition) and of time-varying country-level variables (e. g. social spending); note that the z-variables now have the indices jt because they can (but need not) vary within countries over time. If researchers have a genuine interest in year-level variables (e. g. number of global terror attacks), this model does not work and the model of choice would be a four-level model with individuals nested in country-years, which are cross-classified in countries and years (for more details, see Schmidt-Catran and Fairbrother 2016).

Let us see how our research example plays out with this model. While the models in Table 1 have been fitted to the 2014 wave of the ESS only, the models presented in Table 2 are based on all available ESS data (compare Table 4). Model M3 uses the specification presented above—a three-level model with individuals nested in country-years nested in countries. Model M4 is identical in the fixed part but is a two-level model with individuals nested in countries, i. e. it omits the country-year level. This is a common mistake (compare Schmidt-Catran and Fairbrother 2016, p. 26), as many researchers assume that variables which capture country characteristics are just country-level variables, and do not need a country-year level random effect. As explained above, this is not true if these variables vary over time, as they do in the research example.

Using the pooled data approach and the correct random effects structure (M3), we now find a significant negative effect of social spending, in line with our hypothesis and the result of Immerzeel and Tubergen (2013). Note that the effect of social spending is much weaker than in Table 1 (–0.0137 vs. –0.0465), but it is nonetheless statistically significant. Model M4 demonstrates how a failure to include country-years as a separate level will provide anticonservative SEs. While the point estimates in M3 and M4 are very similar to each other, the z-statistics are much higher in the incorrectly specified model M4 ($|z| = 6.32$ as compared to $|z| = 2.2$). The latter model erroneously treats social spending as an individual-level variable, since it cannot be a country-level variable—because it is not constant within countries.[17]

Model M5 in Table 2 is a two-level model but its random effects structure matches the fixed effects. That is, all country-level variables in the fixed part of the third model have been entered as means, across all years; so they are constant within each country. Consequently, this model should yield correct SEs but it does not benefit from the increase in statistical power. In fact, we gain statistical power at the individual level, where we now have many more observations than in the models in Table 1, but not at the country level.[18] Statistical power at the individual level, however, is typically not scarce with comparative longitudinal survey data and this

[17] Note that this is an oversimplification. Technically, the level at which a variable is measured is not one specific level but it depends on how the variance components of a variable distribute over the levels.

[18] Except for the fact that we now include six additional countries which have been in the ESS at some point but not in the 2014 wave used in Table 1.

Table 2 Random intercept models of church attendance, European Social Survey (ESS) 2002–2014

Variable	M3		M4		M5	
	b/|z|		b/|z|		b/|z|	
Individual-level variables						
Urban vs. rural	0.094	***	0.0939	***	0.0937	***
	40.86	–	40.8	–	40.78	–
Education (in years)	–0.0131	***	–0.0131	***	–0.0134	***
	17.7	–	17.74	–	18.17	–
Subjective income	–0.0272	***	–0.0292	***	–0.0309	***
	7.72	–	8.3	–	8.79	–
Male (ref = female)	–0.2493	***	–0.2499	***	–0.2501	***
	46.5	–	46.56	–	46.6	–
Age	0.0116	***	0.0117	***	0.0117	***
	77.08	–	77.35	–	77.37	–
Country-level variables						
Social spending (% of GDP)	–0.0137	*	–0.0152	***	–0.0385	–
	2.2	–	6.32	–	1.33	–
GDP/capita	–0.0000	–	–0.0000	–	–0.0000	–
	1.01	–	1.83	–	1.17	–
Average urban vs. rural	–0.0542	–	–0.0407	–	0.2738	–
	0.65	–	1.27	–	0.66	–
Average education	–0.0277	–	–0.0198	*	–0.1098	–
	1.3	–	2.32	–	1.37	–
Year FEs						
2004	0.0013	–	0.0009	–	–0.0041	–
	0.05	–	0.09	–	0.41	–
2006	–0.0187	–	–0.0268	–	–0.0473	***
	0.49	–	1.78	–	4.47	–
2008	–0.0158	–	–0.0269	–	–0.0649	***
	0.33	–	1.38	–	6.41	–
2010	–0.0278	–	–0.0388	–	–0.1144	***
	0.52	–	1.8	–	11.22	–
2012	–0.0366	–	–0.048	–	–0.1269	***
	0.6	–	1.9	–	12.43	–
2014	–0.0109	–	–0.0271	–	–0.1171	***
	0.15	–	0.92	–	11.27	–
Constant	2.0519	***	1.9339	***	3.0919	*
	5.15	–	9.91	–	2.55	–
Variance components						
Country	0.339	***	0.3442	***	0.3267	***
Country-year	0.0063	***	–	–	–	–
Individual	1.968	***	1.9729	***	1.9732	***

Table 2 (Continued)

Variable	M3 b/|z|	M4 b/|z|	M5 b/|z|
Statistics			
N (country)	26	26	26
N (country-years)	149	–	–
n (individuals)	277,505	277,505	277,505

See text for explanation of M3–M5
$^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.01$ (two-sided tests). All models are estimated via Restricted Maximum Likelihood. Models based on ESS data 2002–2014 (compare Table 4). Country-level variables in Models M3 and M4 are original variables (yearly record), while in M5 they are averaged across years for each country. Note that Model M4 is incorrectly specified for demonstrational purposes
GDP gross domestic product, *FE* Fixed Effects

is not a recommendation to estimate such models. In fact, the model is presented to motivate the next section. A comparison of the effect of social spending in Models M3 and M5 reveals that it is much larger in the latter, where it is close to the estimate from Table 1 (M2).

3.2 Within-country Estimation of Country-level Effects

The reason for this difference in the effect size is that M2 in Table 1 and M5 in Table 2 are purely cross-sectional estimates; they are the multivariate equivalents of the relationship from the scatter plot in Fig. 2. The estimates from Model M3 in Table 2, which allows country-level variables to vary over time, are identified by two sources of variation: between- and within-country variation and the resulting coefficient is a weighted average of the relationships (Bell et al. 2018; Bell and Jones 2015). Using a variant of Mundlak's (1978) formulation, Fairbrother (2014) demonstrates how comparative longitudinal survey data can be modelled to decompose the total effect into its within- and between-country components. Using the notation from above (but excluding, for ease of presentation, all country-level variables but one), the model can be written like this:

$$y_{jti} = \beta_0 + \sum_{t=1}^{T} \delta_t D_t + \beta_1 x_{1jti} + \ldots + \beta_k x_{kjti} + \gamma^{\mathrm{BE}} \overline{z}_j + \gamma^{\mathrm{WE}} (z_{jt} - \overline{z}_j) + u_j + u_{jt} + e_{jti}$$

The variable $\overline{z}_j$ is the country-level mean of z_{jt} across years; it exhibits only between-country variation, and accordingly γ^{BE} is the between-country effect. The term $(z_{jt} - \overline{z}_j)$ describes the variation of z around the country-specific mean and captures within-country variation; its country-specific mean is zero. The correlation between $(z_{jt} - \overline{z}_j)$ and u_j must be zero. This may sound like a technical detail but it is of utmost importance: The coefficient γ^{WE} provides the within-country effect of z and it cannot suffer from omitted variable bias due to any *time-constant* country-level characteristic because any such unobserved variable would be part of u_j. Thus, the within-effect has an advantage over the between-effect, and the nondecomposed

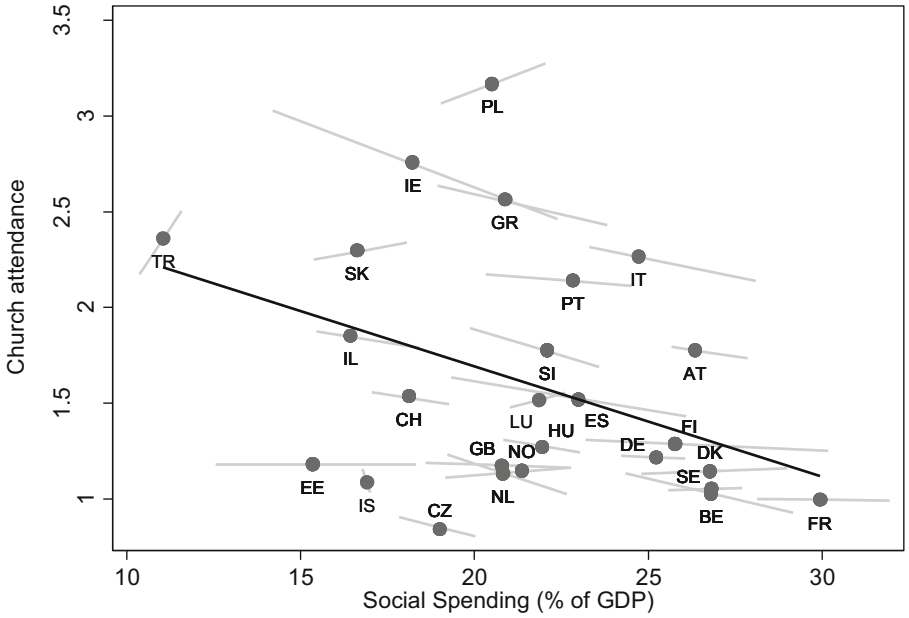

Fig. 3 Bivariate relationships of church attendance and social spending within and between countries. Notes: Based on ESS data 2002–2014 (compare Table 4). The *black line* represents the between-country association, while the *grey lines* represent the associations within the single countries. Compare Fig. 2 for the definition of country codes.

total effect, in terms of the necessary assumptions for unbiasedness (Fairbrother 2014).[19]

In less technical words, the standard interpretation of regression estimates is that "y increases by β units if x increases by one unit". This interpretation clearly implies the notion of change over time. We expect that for any given unit we will observe a change in y because of a change in x. For such a statement to be validly drawn from between-country differences, we assume that the countries in our sample differ only in their observed variables but not in any unobserved (correlated) characteristic. As Gelman (2005, p. 461) puts it "it is a big leap to interpret differences between countries as a potential effect of a change within a country" (Fairbrother 2014, p. 3). It may be a better test to directly investigate change over time within countries.

Figure 3 presents bivariate relationships of social spending and church attendance between countries, as in Fig. 2, but also within each country. The black line represents the between-country association and the grey lines show the within-country relationships. The graph reveals that there are indeed negative relationships between social spending and church attendance in many countries (e. g. Ireland, Italy, Spain,

[19] The idea to identify an effect solely by within-unit variation and thereby to control for any time-constant unobserved variables originates from the analyses of panel data. Readers who want to get a detailed understanding of this may want to read this literature: Allison (2009); Andress et al. (2013, Chap. 4); Bell and Jones (2015).

Table 3 Decomposing country-level effects into within and between components

Variable	M6 b/p		M7 b/p	
Individual-level variables				
Urban vs. rural	0.0940	***	0.0827	***
	0.0000	–	0.0000	–
Education (in years)	–0.0131	***	–0.0131	***
	0.0000	–	0.0000	–
Subjective income	–0.0273	***	–0.0130	***
	0.0000	–	0.0006	–
Male (ref = female)	–0.2494	***	–0.2334	***
	0.0000	–	0.0000	–
Age	0.0116	***	0.0108	***
	0.0000	–	0.0000	–
Country-level variables				
Social Spending (% of GDP) [BE]	–0.0382	–	–0.0139	–
	0.1888	–	0.5394	–
Social Spending (% of GDP) [WE]	–0.0112	–	–0.0081	–
	0.0826	–	0.2338	–
GDP/capita [BE]	–0.0000	–	–0.0000	–
	0.2512	–	0.6884	–
GDP/capita [WE]	–0.0000	–	–0.0000	–
	0.8811	–	0.9078	–
Average urban vs. rural [BE]	0.2796	–	–0.1784	–
	0.5031	–	0.5785	–
Average urban vs. rural [WE]	–0.0485	–	–0.0509	–
	0.5726	–	0.5111	–
Average education [BE]	–0.1109	–	–0.1532	*
	0.1668	–	0.0085	–
Average education [WE]	–0.0139	–	–0.0021	–
	0.5379	–	0.9231	–
Year FEs				
2004	–0.0056	–	–0.0226	–
	0.8401	–	0.3807	–
2006	–0.0428	–	–0.0414	–
	0.2819	–	0.2641	–
2008	–0.0537	–	–0.0514	–
	0.2999	–	0.2794	–
2010	–0.0725	–	–0.0759	–
	0.2109	–	0.1514	–
2012	–0.0899	–	–0.0919	–
	0.1802	–	0.1348	–
2014	–0.0746	–	–0.0815	–
	0.3419	–	0.2566	–

Table 3 (Continued)

Variable	M6 b/p		M7 b/p	
Constant	3.0431	*	3.8998	***
	0.0124	–	0.0000	–
Variance Components				
Country	0.3270	***	0.1663	***
Country-year	0.0062	***	0.0044	***
Individual	1.9680	***	1.9903	***
Statistics				
N (country)	26		23	
N (country-years)	149		128	
n (individuals)	277,505		239,881	

See text for explanation of M6 and M7

*$p<0.05$, **$p<0.01$, ***$p<0.01$ (two-sided tests). All models are estimated via Restricted Maximum Likelihood. Models based on ESS data 2002–2014 (compare Table 4). Model M7 excludes Ireland, Poland and Denmark

GDP gross domestic product, *FE* Fixed Effects, *WE* Within effect, *BE* Between effect

Portugal, Greece, Slovenia) but there are also countries with a positive association (e. g. Norway, Slovakia, Turkey, Luxembourg, Poland) and countries with no apparent relationship (e. g. Estonia, France, Sweden, Great Britain). This casts doubt about the unbiasedness of the cross-sectional analyses presented in Table 1 (M2) and Table 2 (M3). Te Grotenhuis et al. (2015, p. 650) show a similar graph and find a very similar picture. In the example by Te Grotenhuis et al. (2015), it is obvious from the graphical inspection that the vast majority of countries does not show a negative relationship, while in our example one may find an—on average—negative relationship among the countries.

Decomposing country-level effects into their within and between components yields the results presented in Model M6 (Table 3). Within and between-country effects are not identical for any of the four country-level variables, indicating that Model M3 (Table 2), which presented a weighted average of within and between effects, was misleading (see Fairbrother 2014 for a detailed discussion). In all instances, the between effect is much larger than the within effect, indicating that cross-sectional models will often provide overestimated effects due to omitted variable bias. Regarding the effect of social spending, the between effect is –0.038, resembling the effect estimated in the purely cross-sectional Model M2, while the within effect is only –0.011. This coefficient is not significant at the 5%-level but it is close. If the hypothesis is tested one-sided, for which there is a good reason to do because the hypothesis is directed, one could conclude that there is a negative effect of social spending on church attendance; albeit much smaller than a cross-sectional model suggests. So, for now one may conclude that the results of the cross-sectional analyses by Immerzeel and Tubergen (2013), who also tested one-sided, can be replicated by a within-country estimator (but also see the further discussion in the next section).

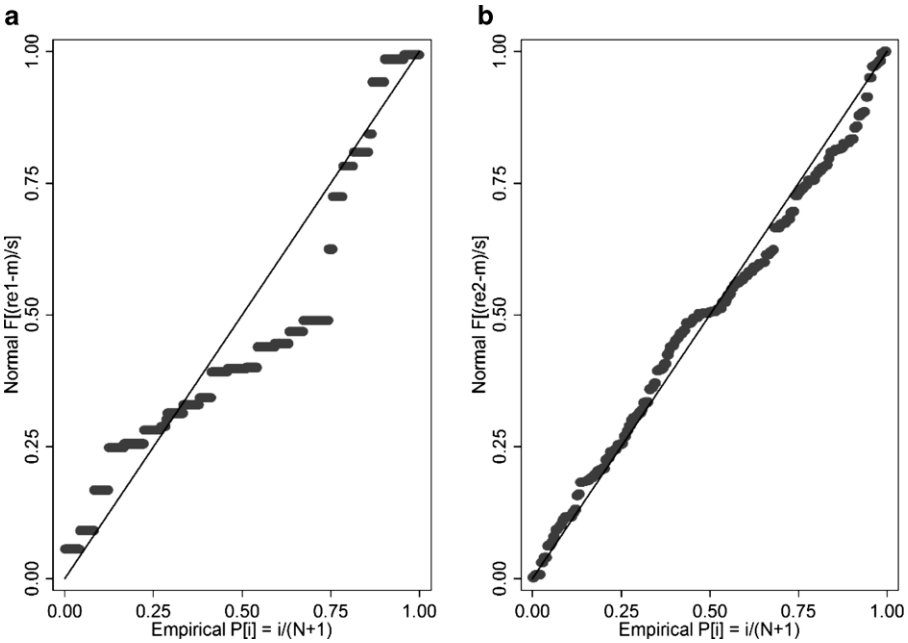

Fig. 4 *P–P* plot of (**a**) country- and (**b**) country-year-level residuals from Model M6 (Table 3). Notes: Based on ESS data 2002–2014 (compare Table 4)

To summarize, by pooling multiple waves of comparative survey data the statistical power can be increased and the option to test hypotheses via within-specifications emerges. From a causal identification standpoint this should be superior to between country estimates which are more prone to omitted variable bias. This is obviously not possible if the variables of interest do not change over time. Similarly, using this technique becomes less useful if the variables of interest change only marginally. In that case, the available variance to identify the effect is small and the estimates will be imprecise. Clearly, the general statistical power and the feasibility of the within-country estimator increases with the number of available survey waves.

4 Diagnostics

Before concluding this article, we briefly discuss diagnostics for multilevel models—specifically, diagnostics for influential cases. There are many statistical tests and graphical inspections that can be used to check for violations of some of the assumptions implicit to the model. Hox (2010, p. 23 ff) provides a very good overview of such tests, and Snijders and Berkhof (2008) discuss many issues in greater and more technical detail. This paper does not have space to discuss regression diagnostics in detail, but that should not be taken as a sign they are not important and

useful. It is valuable to investigate regression diagnostics, particularly through graphical inspection of the residuals at each level. For example, Fig. 4 shows so-called *P–P* plots of the residuals from Model M6 at the country and at the country-year level. These plots allow for a basic visual test of the normality assumption: Perfectly normally distributed residuals form a straight diagonal line. As expected, residuals at the country level (u_j)—with just 26 observations—are not normally distributed, while residuals at the country-year level (u_{jt}), with 139 observations, are close to a normal distribution. In principle, violations of the normality assumption can result in biased SEs and require some caution with respect to statistical inference, though simulations reported by Bell et al. (2019) suggest that such biases are in practice quite modest.

Cook's Distance (Cook's D, for short) is a measure describing the influence of single observations on all estimated coefficients (Cook 1977). In the context of multilevel models, it can be applied to the random and the fixed part separately (Snijders and Berkhof 2008, p. 157 ff.):

$$C_j^F = \frac{1}{r}\left(\widehat{\beta} - \widehat{\beta}_{(-j)}\right)' \widehat{S}^{-1}_{F(-j)} \left(\widehat{\beta} - \widehat{\beta}_{(-j)}\right), \text{ for the fixed part and}$$

$$C_j^R = \frac{1}{p}\left(\widehat{\eta} - \widehat{\eta}_{(-j)}\right)' \widehat{S}^{-1}_{R(-j)} \left(\widehat{\eta} - \widehat{\eta}_{(-j)}\right), \text{ for the random part;}$$

where $\widehat{\beta}$ and $\widehat{\eta}$ are *vectors* of parameter estimates from the fixed and the random part, respectively, and $\widehat{\beta}_{(-j)}$ and $\widehat{\eta}_{(-j)}$ are the same estimates when country j is left out from the sample. Finally, $\widehat{S}_{F(-j)}$ and $\widehat{S}_{R(-j)}$ are the estimated covariance matrices of the fixed and random part and r and p are the numbers of parameters estimated in the fixed and random part. Cook's D can be interpreted as the standardized average squared difference in parameter estimates with and without country j (Van der Meer et al. 2010, p. 175). The total Cook's D measure for the model equals the weighted average of Cook's D for the random and the fixed part:

$$C_j = \frac{1}{r+p}\left(rC_j^F + pC_j^R\right).$$

Since hypotheses are most often about the fixed part of the model, researchers may want to examine the single components rather than the total measure. And with very few countries it is entirely possible that every country will appear to be influential.

That of course depends on the definition of "too influential". Belsley et al. (1980, p. 28) propose the cut-off value $4/n$ for Cook's D. Table 5 (appendix) presents Cook's D for the fixed part of Model M6 and 19 out of 26 countries are deemed too influential if we follow the proposal of Belsley et al. (1980).[20] Obviously, this is not very helpful because the exclusion of 19 countries is not an option. Nevertheless, the Cook's D measure indicates which of the countries has the strongest influence

[20] For two-level models, Stata users can use the mlt ado-package to calculate Cook's D and DFBETAs (Möhring and Schmidt 2013). In the online appendix we provide a syntax for three-level models which is very general and can be easily adapted to researchers' own applications.

on the sum of all (fixed) parameter estimates. In the example data the by far most influential countries are Ireland (Cook's D = 5.12) and Israel (Cook's D = 4.84, where the next highest-ranked country has a value of about 2.6). Looking at Fig. 3, one may wonder why Israel (IL) appears as influential, given its position in the scatter plot; recall though that Cook's D is based on the *sum of all parameter estimates* from a multivariate model, not on bivariate relationships.

Nevertheless, to decide how robust the conclusion about the social spending effect is, Cook's D may not be the best measure. After all, Israel does not appear to be a suspicious case in Fig. 3, neither regarding the between- nor the within-country relationship. DFBETA is a measure that describes the influence of a single unit on a selected coefficient (Belsley et al. 1980, p. 13) and can be applied in the context of multilevel models as well (Van der Meer et al. 2010, p. 175):

$$\text{DFBETA}_{zj} = \frac{\widehat{\beta}_z - \widehat{\beta}_{z(-j)}}{\text{SE}\left(\widehat{\beta}_{z(-j)}\right)},$$

where $\widehat{\beta}_z - \widehat{\beta}_{z(-j)}$ is the difference between the effects of variable z with and without country j.[21] This difference is divided by the SE of the effect in the model without country j. DFBETAs can be understood as the standardized difference between the coefficients with and without unit j. For DFBETAs, Belsley et al. (1980, p. 28) propose the cut-off value $2/\sqrt{n}$. Table 5 in the appendix presents DFBETAs for the within effect of social spending and identifies three influential cases: Denmark, Ireland and Poland, with Ireland having a strong negative impact on the estimates (DFBETA = –1.85) and Denmark and Poland having positive influences (0.54 and 0.76, respectively). Again, not all of these countries seem suspicious from inspecting Fig. 3. Given the country-specific relationships presented in Fig. 3, it is no wonder that Ireland has a strong negative impact and that Poland has a positive effect on the estimates but Denmark seems to be a rather inconspicuous case. This is precisely the reason why graphical inspections of bivariate relationships alone are not sufficient to identify influential cases (Van der Meer 2010, p. 175).

The blind exclusion of countries, because they exceed some cut-off value in an outlier statistic, is not a useful strategy; but paying attention to these cases certainly is. One may argue that these outliers are valuable candidates for case studies and/or hint at the need for better theories. Since the space in this paper is limited, we simply present estimates without these three influential cases (Ireland, Poland and Denmark) in Model M7 (Table 3). Focusing only on the effect of interest, Model M7 yields a smaller within effect of social spending than Model M6, and the *p*-value for the coefficient on social spending has increased substantially.

[21] DFBETAs can of course also be calculated for individual-level variables (x) but in the context of multilevel modeling its application to country-level variables (z) is typically of interest.

5 Conclusions

We will not attempt to settle the debate between Immerzeel and Tubergen's (2013) argument that social spending has a negative effect on church attendance and the objection by Te Grotenhuis et al. (2015) that this result does not stand up when tested longitudinally. Instead, the purpose of this exercise has been to demonstrate how sensitive results from multilevel models with comparative survey data can be to various decisions taken during the research process, and to suggest useful ways of thinking about that sensitivity.

This paper has addressed a selection of issues, but there are others it has ignored. The research example in this paper used a linear multilevel model, and while all issues are also relevant for nonlinear models, the nonlinear case presents some additional challenges (see Bryan and Jenkins 2016). We also did not address in great detail the estimation of cross-level interactions and random slopes, both of which are important topics (see Bell et al. 2019; Elff et al. 2016; Giesselmann and Schmidt-Catran in press). Finally, we also did not address any issues of model building. For the example analysis, we simply took the model from Te Grotenhuis et al. (2015). Particularly where degrees of freedom are limited, researchers need to choose what variables to include very carefully, on both theoretical and empirical grounds.

Comparative survey data are characterized by a small number of higher-level units (countries) which are not random samples. This presents researchers with several challenges, including questions about whether inferential statistics are useful at all, what the appropriate estimation method is, and whether estimates are sensitive to single countries. There is also a risk of omitted variable bias, or the inability to include a full complement of variables and/or random slopes. While inference about country-level effects must be viewed as conditional on the observed sample, inferential statistics are, from our view, still useful in the context of multilevel models fitted to comparative survey data. With small samples at the country level, researchers would do well to test the robustness of their findings to the choice of different estimation methods. While statistical power at the country level is typically scarce, the contrary is true for the individual level, where observation numbers are typically very large. Particularly at this level, researchers should always consider the practical size of the effects in addition to their levels of significance.

The issue of omitted variables can—to some extent—be addressed by employing within-country estimators, though this requires observing sufficient change over time in the variable of interest and to have a decent number of waves that can be pooled. Thus, not every research question can be tested with these methods. Obviously, such an estimator does only control for time-constant omitted variables but it can still suffer from omitted variables if these too vary over time.

Appendix

Table 4 Sample sizes of example data—European Social Survey (ESS) rounds 1 to 7

Country	Year							Total
	2002	2004	2006	2008	2010	2012	2014	
AT	2138	2154	2288	–	–	–	1778	8358
BE	1725	1742	1781	1747	1662	1863	1759	12,279
CH	2002	2111	1787	1785	1473	1477	1508	12,143
CZ	1224	2525	–	1934	2279	1737	1943	11,642
DE	2849	2723	2803	2688	2986	2920	3001	19,970
DK	1461	1451	1443	1584	1557	1621	1487	10,604
EE	–	1970	1471	1596	1784	2358	2016	11,195
ES	1532	1589	1743	2481	1834	1838	1849	12,866
FI	1972	1997	1880	2181	1852	2169	2064	14,115
FR	–	–	1965	2039	1714	1952	1895	9565
GB	2007	1862	2342	2300	2352	2222	2221	15,306
GR	2515	2388	–	2039	2667	–	–	9609
HU	1672	1471	1481	1511	1548	1937	1645	11,265
IE	1930	2195	1586	1743	2514	2582	2316	14,866
IL	2289	–	–	2264	2020	2353	2460	11,386
IS	–	551	–	–	–	720	–	1271
IT	1160	1484	–	–	–	856	–	3500
LU	1403	1567	–	–	–	–	–	2970
NL	2309	1847	1870	1748	1784	1825	1881	13,264
NO	2027	1754	1744	1541	1540	1614	1432	11,652
PL	2071	1695	1687	1595	1703	1840	1563	12,154
PT	1456	1989	2072	2237	2003	2066	1239	13,062
SE	1979	1924	1903	1811	1488	1812	1761	12,678
SI	1487	1369	1433	1234	1352	1233	1210	9318
SK	–	1373	1649	1725	1790	1784	–	8321
TR	–	1805	–	2341	–	–	–	4146
Total	39,208	43,536	34,928	42,124	39,902	40,779	37,028	277,505

Source: ESS data 2002–2014, obtained from the cumulative data wizard on 20th September 2017
Note: For definition of the country codes compare Fig. 2.

Table 5 Cook's D of fixed part and DFBETAs of within-effect of social spending from Model M6

Country	Cook's D	DFBETAs
AT	0.2983*	0.0025
BE	0.4002*	–0.1199
CH	0.0952	–0.0782
CZ	0.0786	0.0474
DE	0.5876*	–0.2020
DK	0.2882*	0.5436*
EE	0.6936*	0.1051
ES	1.7909*	–0.0312
FI	0.6411*	0.2576
FR	0.3415*	0.1015
GB	1.2517*	0.0654
GR	0.6241*	–0.0859
HU	0.1458	0.1163
IE	5.1165*	–1.8460*
IL	4.8394*	–0.3309
IS	0.0062	–0.0238
IT	0.0525	0.0137
LU	0.0316	0.1129
NL	0.1516	–0.0785
NO	0.2899*	0.0436
PL	0.4649*	0.7625*
PT	2.6162*	0.0936
SE	0.4284*	–0.2783
SI	0.6803*	0.1920
SK	0.5251*	0.0055
TR	2.1563*	0.2088

* cut-off value (Cook's D = 0.1538, DFBETAs = 0.3922) exceeded
Note: For definition of the country codes compare Fig. 2.

References

Allison, Paul D. 2009. *Fixed effects regression models*. Thousand Oaks: SAGE.

Andress, Hans-Jürgen, Katrin Golsch and Alexander W. Schmidt. 2013. *Applied panel data analysis for economic and social surveys*. Springer: Berlin, Heidelberg.

Arceneaux, Kevin, and Gregory A. Huber. 2007. What to do (and not do) with multicollinearity in state politics research. *State Politics & Policy Quarterly* 7:81–101.

Babones, Salvatore J. 2013. *Methods for quantitative macro-comparative research*. London: Sage.

Barr, Dale J., Roger Levy, Christoph Scheepers and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68:255–278.

Beck, Nathaniel. 2001. Time-series–cross-section data: What have we learned in the past few years? *Annual Review of Political Science* 4:271–293.

Bell, Andrew J., and Kelvyn Jones. 2015. Explaining fixed effects: Random effects modeling of time-series cross-sectional and data panel data. *Political Science Research and Methods* 3:133–153.

Bell, Andrew J., Kelvyn Jones and Malcolm Fairbrother. 2018. Understanding and misunderstanding Group mean centering: A commentary on Kelley et al.'s dangerous practice. *Quality & Quantity* 52:2031–2046.

Bell, Andrew, Malcolm Fairbrother and Kelvyn Jones. 2019. Fixed and random effects models: Making an informed choice. *Quality & Quantity* 53:1051–1074.
Bell, Bethany A., Grant B. Morgan, Jason A. Schoeneberger, Jeffrey D. Kromrey and John M. Ferron. 2014. How low can you go? *Methodology* 10:1–11.
Belsley, David A., Edwin Kuh and Roy E. Welsch. 1980. *Regression diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley: New York.
Bowers, Jake, and Katherine W. Drake. 2005. EDA for HLM: Visualization when probabilistic inference fails. *Political Analysis* 13:301–326.
Brännström, Lars. 2008. Making their mark: The effects of neighbourhood and upper secondary school on educational achievement. *European Sociological Review* 24:463–478.
Browne, William J., and David Draper. 2000. Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics* 15:391–420.
Bryan, Mark L., and Stephen P. Jenkins. 2016. Multilevel modelling of country effects: A cautionary tale. *European Sociological Review* 32:3–22.
Cook, R. Dennis. 1977. Detection of influential observation in linear regression. *Technometrics* 19:15–18.
Draper, David. 2008. Bayesian multilevel analysis and MCMC. In *Handbook of multilevel analysis*, eds. Jan De Leeuw and Erik Meijer, 77–139. Springer: New York.
Ebbinghaus, Bernhard. 2005. When less is more: selection problems in large-N and small-N cross-national comparisons. *International Sociology* 20:133–152.
Elff, Martin, Jan P. Heisig, Merlin Schaeffer and Susumu Shikano. 2016. No need to turn Bayesian in multilevel analysis with few clusters: How frequentist methods provide unbiased estimates and accurate inference. *Open Science Framework* (osf.io/fkn3u).
Elwert, Felix. 2013. Graphical causal models. In *Handbook of causal analysis for social research,* ed. Morgan, Stephen L., 245–273. Springer Netherlands: Dordrecht.
Enders, Craig K., and Davood Tofighi. 2007. Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods* 12:121–138.
European Social Survey Cumulative File, ESS 1–7 (2016). Data file edition 1.0. NSD – Norwegian Centre for Research Data, Norway – Data Archive and distributor of ESS data for ESS ERIC.
Fairbrother, Malcolm. 2013. Rich people, poor people, and environmental concern: Evidence across nations and time. *European Sociological Review* 29:910–922.
Fairbrother, Malcolm. 2014. Two multilevel modeling techniques for analyzing comparative longitudinal survey datasets. *Political Science Research and Methods* 2:119–140.
Fairbrother, Malcolm. 2016. Trust and public support for environmental protection in diverse national contexts. *Sociological Science* 3: 359–382.
Finseraas, H. 2012. Poverty, ethnic minorities among the poor, and preferences for redistribution in European regions. *Journal of European Social Policy* 22:164–180.
Fontaine, Johnny R.J. 2015. Traditional and multilevel approaches in cross-cultural research: An integration of methodological frameworks. In *Multilevel analysis of individuals and cultures,* eds. Van de Vijver, Fons J.R., Dianne A. Van Hemert and Ype H. Poortinga, 65–93. New York: Psychology Press.
Gelman, Andrew. 2005. Two-stage regression and multilevel modeling: A commentary. *Political Analysis* 13:459–61.
Gelman, Andrew. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 1:72–91.
Gelman, Andrew, and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.
Giesselmann, Marco and Alexander W. Schmidt-Catran. 2018. Getting the within estimator of cross-level interactions in multilevel models with pooled cross-sections: Why country dummies (sometimes) do not do the job. *Sociological Methodology*. https://doi.org/10.1177/0081175018809150
Giger, Nathalie. 2012. Is social policy retrenchment unpopular? How welfare reforms affect government popularity. *European Sociological Review* 28:691–700.
Goerres, Achim, Markus B. Siewert, and Claudius Wagemann. 2019. Internationally comparative research designs in the social sciences: Fundamental issues, case selection logics, and research limitations. In Cross-national comparative research – analytical strategies, results and explanations. Sonderheft *Kölner Zeitschrift für Soziologie und Sozialpsychologie*. Eds. Hans-Jürgen Andreß, Detlef Fetchenhauer and Heiner Meulemann. Wiesbaden: Springer VS. https://doi.org/10.1007/s11577-019-00600-2.
Goldthorpe, John H. 1997. Current issues in comparative macrosociology: A debate on methodological issues. *Comparative Social Research* 16:1–26.

Te Grotenhuis, Manfred, Marijn Scholte, Nan Dirk de Graaf and Ben Pelzer. 2015. The between and within effects of social security on church attendance in Europe 1980–1998: The danger of testing hypotheses cross-nationally. *European Sociological Review* 31:643–654.
Heisig, Jan P., Merlin Schaeffer and Johannes Giesecke. 2017. The costs of simplicity: Why multilevel models may benefit from accounting for cross-cluster differences in the effects of controls. *American Sociological Review* 82:796–827.
Heisig, Jan Paul, and Merlin Schaeffer. 2019. Why You Should Always Include a Random Slope for the Lower-Level Variable Involved in a Cross-Level Interaction. *European Sociological Review* 35:258–279.
Hox, Joop J. 2010. *Multilevel analysis: Techniques and applications, 2nd Edition*. New York: Routledge.
Immerzeel, Tim, and Frank Van Tubergen. 2013. Religion as reassurance? Testing the insecurity theory in 26 European countries. *European Sociological Review* 29:359–372.
Jackman, Simon D. 2009. Bayesian analysis for the social sciences. New York: John Wiley.
Jaeger, Mads M. 2013. The effect of macroeconomic and social conditions on the demand for redistribution: A pseudo panel approach. *Journal of European Social Policy* 23:149–163.
Kim, Jee-Seon, and Edward W. Frees. 2006. Omitted variables in multilevel models. *Psychometrika* 71:659–690.
Maas, Cora J. M., and Joop J. Hox. 2005. Sufficient sample sizes for multilevel modeling. *methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 1:86–92.
Van der Meer, Tom, Manfred Te Grotenhuis and Ben Pelzer. 2010. Influential cases in multilevel modeling: A methodological comment. *American Sociological Review* 75:173–178.
Mewes, Jan. 2014. Gen (d) eralized trust: women, work, and trust in strangers. *European Sociological Review* 30:373–386.
Möhring, Katja, and Alexander W. Schmidt. 2013. MLT: Stata module to provide multilevel tools (Statistical Software Components S457577). Boston, MA: Boston College Department of Economics.
Mundlak, Yair. 1978. Pooling of time-series and cross-section data. *Econometrica* 46:69–85.
Patterson, H. Desmond, and Robin Thompson. 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika* 58:545–554.
Rabe-Hesketh, Sophia, and Anders Skrondal. 2012. *Multilevel and longitudinal modeling using stata, 3rd Edition*. College Station, TX: Stata Press.
Schmidt-Catran, Alexander W. 2016. Economic inequality and demand for redistribution: Cross-sectional and longitudinal evidence. *Socio-Economic Review* 14:119–140.
Schmidt-Catran, Alexander W., and Malcom Fairbrother. 2016. The random effects in multilevel models: Getting them wrong and getting them right. *European Sociological Review* 32:23–38.
Semyonov, Moshe, Rebeca Raijman and Anastasia Gorodzeisky. 2006. The rise of anti-foreigner sentiment in European societies, 1988–2000. *American Sociological Review* 71:426–449.
Snijders, Tom A.B., and Johannes Berkhof. 2008. Diagnostic checks for multilevel models. In *Handbook of multilevel analysis*, eds. Jan De Leeuw and Erik Meijer, 457–514. New York: Springer.
Snijders, Tom A.B., and Roel J. Bosker. 2012. *Multilevel analysis: An introduction to basic and advanced multilevel modelling, 2nd Edition.* London: Sage.
Stadelmann-Steffen, Isabelle. 2012. Education policy and educational inequality—evidence from the Swiss laboratory. *European Sociological Review* 28:379–393.
Stegmueller, Daniel. 2013. How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science* 57:748–761.
Stegmueller, Daniel, Peer Scheepers, Sigrid Roßteutscher and Eelke de Jong. 2012. Support for redistribution in western Europe: Assessing the role of religion. *European Sociological Review* 28:482–497.
Sutton, John R. 2012. Imprisonment and opportunity structures: A Bayesian hierarchical analysis. *European Sociological Review* 28:12–27.
Van Erp, Sara, Joris Mulder and Daniel L. Oberski. 2017. Prior sensitivity analysis in default Bayesian structural equation modeling. *Psychological Methods* 23:363–388.
Wilkes, Rima, Neil Guppy and Lily Farris. 2007. Right-wing parties and anti-foreigner sentiment in Europe. *American Sociological Review* 72:831–840.
Wooldridge, Jeffrey M. 2013. *Introductory econometrics: A modern approach*. Mason, OH: South-West Cengage Learning.
Wulfgramm, M. 2014. Life satisfaction effects of unemployment in Europe: The moderating influence of labour market policy. *Journal of European Social Policy* 24:258–272.

Alexander W. Schmidt-Catran 1983, Dr. rer.pol., Professor for sociology with a focus on quantitative methods, Goethe-University Frankfurt. Areas of research: welfare states, migration, public opinion, statistical methods. Publications: The random effects in multilevel models: Getting them wrong and getting them right. European Sociological Review 32, 2016 (together with M. Fairbrother); Applied panel data analysis for economic and social surveys. Berlin, Heidelberg 2013 (together with H.-J. Andreß and K. Golsch); Immigration and welfare support in Germany. American Sociological Review 81, 2016 (together with D. Spies).

Malcolm Fairbrother 1975, PhD, Professor of Sociology, Umeå University, and researcher, Institute for Futures Studies, Stockholm. Areas of research: political sociology, environment/nature-society relations, globalization, trust, social science methodology. Publications: When will people pay to pollute? Environmental taxes, political trust, and experimental evidence from Britain. British Journal of Political Science 46, 2017; Economists, capitalists, and the making of globalization: North American free trade in comparative-historical perspective. American Journal of Sociology 119, 2014; Two multilevel modeling techniques for analyzing comparative longitudinal survey datasets. Political Science Research and Methods 47, 2014.

Hans-Jürgen Andreß 1952, Dr. phil., Professor for Empirical Social and Economic Research. Areas of research: social research, statistics and multivariate methods, social inequality, labor market research, social and family policy. Publications: Is material deprivation decreasing in Germany? A trend analysis using PASS data from 2006 to 2013. Journal for Labour Market Research 52, 2018; The economic consequences of divorce in Germany: What has changed since the turn of the millennium? Comparative Population Studies 40, 2015 (with M. Bröckel); Applied panel data analysis for economic and social surveys. Berlin 2013 (together with K. Golsch).