Heisig, Jan Paul; Schaeffer, Merlin; Giesecke, Johannes

The Costs of Simplicity: Why Multilevel Models May Benefit from Accounting for Cross-Cluster Differences in the Effects of Controls

American Sociological Review

AMERICAN SOCIOLOGICAL ASSOCIATION

American Sociological Review
2017, Vol. 82(4) 796–827

DOI: 10.1177/0003122417717901
journals.sagepub.com/home/asr

SAGE

The Costs of Simplicity: Why Multilevel Models May Benefit from Accounting for Cross-Cluster Differences in the Effects of Controls

Jan Paul Heisig,[a] Merlin Schaeffer,[b] and Johannes Giesecke[c]

Abstract

Context effects, where a characteristic of an upper-level unit or cluster (e.g., a country) affects outcomes and relationships at a lower level (e.g., that of the individual), are a primary object of sociological inquiry. In recent years, sociologists have increasingly analyzed such effects using quantitative multilevel modeling. Our review of multilevel studies in leading sociology journals shows that most assume the effects of lower-level control variables to be invariant across clusters, an assumption that is often implausible. Comparing mixed-effects (random-intercept and slope) models, cluster-robust pooled OLS, and two-step approaches, we find that erroneously assuming invariant coefficients reduces the precision of estimated context effects. Semi-formal reasoning and Monte Carlo simulations indicate that loss of precision is largest when there is pronounced cross-cluster heterogeneity in the magnitude of coefficients, when there are marked compositional differences among clusters, and when the number of clusters is small. Although these findings suggest that practitioners should fit more flexible models, illustrative analyses of European Social Survey data indicate that maximally flexible mixed-effects models do not perform well in real-life settings. We discuss the need to balance parsimony and flexibility, and we demonstrate the encouraging performance of one prominent approach for reducing model complexity.

Keywords

multilevel modeling, comparative research, cluster-robust standard errors, hierarchical data, context effects

A primary object of sociological inquiry is how social environments—ranging from the family and the neighborhood to the state—shape human action. An enthusiasm for such context effects motivates many of the greatest examples of sociological thought—from Durkheim's (1897) classical investigation of anomie in *Le Suicide* to Sampson's (2013) recent portrayal of the *Great American City*.

In recent decades, scholars have increasingly examined context effects using quantitative statistical analysis, thanks in part to a growing availability of suitable data. Such analyses

[a]WZB Berlin Social Science Center
[b]University of Cologne
[c]Humboldt University Berlin

Corresponding Author:
Merlin Schaeffer, University of Cologne, Albertus-Magnus-Platz, D-50923 Köln, Germany
E-mail: merlin.schaeffer@uni-koeln.de

pose statistical challenges because observations belonging to the same contextual unit or cluster (e.g., a school, census district, or country) tend to be more similar than two randomly chosen observations, violating the independence assumption of traditional regression analysis. Statisticians and practitioners in sociology and other disciplines have developed several distinct approaches for addressing this complication, which we subsume under the term "multilevel modeling techniques." In sociology, the most prevalent approach is to use mixed-effects multilevel or hierarchical models with random intercepts and random slopes. Mixed-effects models are powerful tools for analyzing multilevel data; but they are not the only ones. Economists tend to favor models with cluster-robust standard errors, whereas some scholars, especially in political science, advocate two-step approaches. A first goal of our article is to introduce applied sociologists to these alternative techniques and to highlight their strengths and weaknesses.

Our second and more important goal is to explore the consequences of a common assumption in applied research: that the coefficients of lower-level control variables do not vary across clusters. Our review of all articles published in three leading generalist sociology journals between 2011 and 2014 indicates that this "invariant coefficients assumption" (ICA) is nearly ubiquitous in quantitative multilevel studies. Yet such an assumption is often highly unrealistic, as we will demonstrate using illustrative analyses of European Social Survey (ESS) data.

Adding to a small but growing literature on the shortcomings of current multilevel modeling practice (e.g., Bryan and Jenkins 2016; Schmidt-Catran and Fairbrother 2016), we explore the ramifications of assuming invariant coefficients through semi-formal reasoning, Monte Carlo simulations, and empirical analyses. Drawing on an analogy to classic omitted variable bias, our semi-formal argument suggests that neglecting differences in the effects of control variables adds noise to cluster-level relations of interest. Coefficient estimates of context effects thus become less precise. For the same reason, allowing effects of controls to vary across clusters (e.g., by adding random slopes to a mixed-effects model) can result in more precise estimates of context effects. Increased precision means that the estimates produced by individual studies tend to be closer to the corresponding true values. Put differently, the risk that a single study severely misrepresents a context effect will be minimized. More generally, repeated studies will produce more similar results and standard errors will be smaller, raising statistical power (i.e., the probability of rejecting the null hypothesis of no effect when it is false). The potential loss of accuracy associated with the invariant coefficients assumption should be a major concern in multilevel applications: because there are often very few clusters, estimates of contextual effects tend to be quite uncertain. Using an unnecessarily imprecise estimator exacerbates this problem.

We present extensive Monte Carlo simulations to explore this issue under controlled and ideal conditions. We consider several prominent multilevel estimation strategies (i.e., mixed-effects models, cluster-robust pooled OLS, and two-step approaches) and compare three basic scenarios that are typical of applied research. The first two resemble cross-national comparisons with a small number of clusters (15 and 25) and large samples at the lower level (between 600 and 2,000 units per cluster). The third scenario is more typical of applications with other types of contextual units, such as cities or neighborhoods (50 clusters with 70 to 130 units per cluster). To assess the consequences of the ICA, we introduce cluster-specific coefficients for up to five lower-level variables, with effect heterogeneity being unrelated to the contextual factor of interest. Given this benign form of heterogeneity, the ICA does not lead to biased parameter estimates. In line with our argument, however, it does result in a loss of precision relative to more flexible specifications. The severity of the problem depends on the extent of (neglected) heterogeneity, the number of clusters, and the extent

of compositional differences among them (with respect to lower-level predictors).

The simulation results suggest that the ICA entails an avoidable loss of precision and that practitioners have much to gain from using more flexible specifications. To investigate this possibility under realistic conditions, we conduct illustrative multilevel analyses using data on 28 countries from the European Social Survey. Our analyses show that the maximally flexible mixed-effects specification with random slopes for all lower-level controls and no constraints on the correlations among them typically performs worse than the invariant specification. Thus, practitioners should not respond to our simulation results by blindly estimating a model that allows all coefficients to vary. Rather, our overall conclusion is that it will often be possible to improve the precision of estimated context effects by finding a random-effects specification that captures the most important patterns of cross-cluster heterogeneity, yet remains parsimonious enough to be estimable using a small upper-level sample.

The task of finding the optimal random-effects specification is essentially one of model selection, so it is possible to draw on a well-established toolbox. In our applications, we use a slight modification of Bates, Kliegl, Vasishth, and Baayen's (2015) procedure for simplifying complex random-effects structures. In addition to classic indicators such as information criteria, Bates, Kliegl, and colleagues (2015) use principal component analysis to identify overly complex random-effects structures that are not supported by the data. In our applications, the method yields models that compare favorably with the invariant specification that predominates in applied research. We conclude that the method is a promising approach for simplifying random-effects structures in typical sociological applications. However, we note that alternative selection strategies that are more specifically tailored toward the identification of context effects might perform even better.

DIFFERENT APPROACHES TO HANDLING MULTILEVEL DATA

Before we explore the implications of ignoring cross-cluster variation in the effects of lower-level control variables, we introduce the statistical challenges of multilevel analysis and the three estimation approaches that we compare in this study: cluster-robust ordinary least squares (OLS), mixed-effects (ME), and two-step estimation. A simple linear model is a useful starting point:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i. \quad (1)$$

y_i is the value of the outcome variable for unit i. Throughout the article, we focus on the case where y_i is metric (or continuous). Equation 1 specifies y_i as a linear combination of a constant β_0, i's values on covariates x_{1i} to x_{ki}, multiplied by their respective coefficients β_1 to β_k, and an unobservable error term ε_i.

The primary goal in regression analysis usually is to estimate parameters β_0 to β_k. The most popular approach for doing so is ordinary least squares (OLS) estimation. However, multilevel data typically violate the fundamental OLS assumption that the error term ε_i is uncorrelated across observations. Observations from the same country or school likely share unobserved (contextual and individual) attributes that render them more similar than observations from different clusters. OLS estimation remains unbiased when the independence assumption does not hold (Kloek 1981; Moulton 1986). However, it is no longer the most precise (or most efficient) estimator; other estimators can produce more reliable estimates that tend to be closer to the true parameter values. Second, the conventional estimator of OLS standard errors is no longer appropriate, resulting in invalid—and usually anticonservative—statistical inference (Moulton 1986, 1990). That is, confidence intervals based on conventional OLS standard errors will be too narrow and p-values

too small, leading to over-rejection of the null hypothesis of no effect.

Cluster-Robust Pooled OLS

Cluster-robust pooled OLS sticks with OLS for estimating slope parameters and only corrects standard error estimates to achieve accurate statistical inference (for an extensive treatment, see Cameron and Miller 2015). Cluster-robust standard errors are an extension of White's (1980) heteroscedasticity-consistent standard errors to clustered settings (Rogers 1993; Williams 2000). They remain consistent given any kind of between- and within-cluster heteroscedasticity and autocorrelation (Wooldridge 2003).

In practice, cluster-robust standard errors tend to be larger than conventional OLS standard errors, and often substantially so. The difference between conventional and cluster-corrected standard errors increases with average cluster size and the extent of imbalance (i.e., variation in cluster sizes). Moreover, it is larger the less a predictor varies within clusters (Kloek 1981; Moulton 1986, 1990). Hence, failure to correct for clustering is particularly consequential for statistical inference on contextual variables, which do not vary within clusters.

Cluster-robust standard errors are a flexible and easy-to-use means of accounting for clustering that has been generalized to many other estimators, including nonlinear models such as probit and logit. Unfortunately, conventional cluster-robust standard errors require a sufficient number of clusters to be fully accurate; simulation studies suggest they are too small when the cluster-level sample is below 50 (e.g., Kézdi 2004; MacKinnon and Webb 2014). Hence, they will be inappropriate in many sociological applications, and researchers should consider more recently developed small-sample corrections (e.g., Esarey and Menger 2015; Imbens and Kolesár 2016).

Mixed-Effects Models

Mixed-effects models allow the analyst to explicitly model multilevel structures, which can increase the precision of point estimates relative to pooled OLS (Gelman and Hill 2007; Lindley and Smith 1972). Mixed-effects multilevel models are usually estimated via (restricted) maximum likelihood, but Bayesian estimation is an important alternative (Gelman and Hill 2007). To illustrate the approach, reconsider the model in Equation 1, adding the g subscript to index clusters ($g = 1, \ldots, G$):

$$y_{ig} = \beta_{0g} + \beta_{1g}x_{1ig} + \ldots + \beta_{kg}x_{kig} + \varepsilon_{ig}. \quad (2)$$

The basic idea of mixed-effects modeling is to assume that one or several of the β parameters in Equation 2 vary randomly over the G clusters. The simplest model, a random-intercept model, assumes that only β_{0g} varies due to a cluster-level random effect v_{0g}. In a direct-context-effect (DCE) model, β_{0g} also depends on at least one cluster-level covariate z_g, yielding the following equation:

$$\beta_{0g} = \gamma_{00} + \gamma_{01}z_g + v_{0g}. \quad (3)$$

γ_{00} is the overall intercept and γ_{01} coefficient on z_g, a direct context effect. The label "mixed-effects model" is due to the fact that these models include both fixed effects (coefficients) such as γ_{00} and γ_{01} and random effects such as v_{0g}.

Mixed-effects models can further incorporate cluster differences in the coefficients of lower-level variables by incorporating random slopes and cross-level interactions. As an example, consider the following model for β_{1g}, the slope of x_{1ig}:

$$\beta_{1g} = \gamma_{10} + \gamma_{11}z_g + v_{1g}. \quad (4)$$

According to this equation, β_{1g} depends on z_g and a random effect v_{1g}. Substitution of Equations 3 and 4 into 2 yields the following:

$$\begin{aligned} y_{ig} &= (\gamma_{00} + \gamma_{01}z_g + v_{0g}) + (\gamma_{10} + \gamma_{11}z_g + v_{1g})x_{1ig} \\ &\quad + \ldots + \beta_{kg}x_{kig} + \varepsilon_{ig} \\ &= \gamma_{00} + \gamma_{01}z_g + \gamma_{10}x_{1ig} + \gamma_{11}z_g x_{1ig} + \ldots + \\ &\quad \beta_{kg}x_{kig} + (v_{0g} + v_{1g}x_{1ig} + \varepsilon_{ig}). \end{aligned} \quad (5)$$

This equation now includes two random effects, v_{0g} and v_{1g}, which account for within-cluster

similarities that remain after accounting for lower- and cluster-level predictors. These random effects are usually assumed to follow a multivariate normal distribution. $z_g x_{1ig}$ is a cross-level interaction (CLI) between a cluster-level and a lower-level covariate. It is possible to specify more than one slope as random and to include additional cluster-level characteristics.

Mixed-effects models borrow strength by partially pooling observations across clusters, shrinking estimates for unreliable clusters (e.g., with small samples) toward the cluster-average intercept or slopes (Gelman and Hill 2007; Raudenbush and Bryk 2002). A potential weakness of mixed-effects models is the assumption of (multivariate) normally distributed random effects, although simulation studies suggest they are rather robust to deviations from normality (Maas and Hox 2004).

Two-Step Estimation

Two-step estimation is perhaps the most intuitive approach to multilevel modeling. The basic idea is to estimate G cluster-specific regressions in a first step:

$$\begin{aligned} y_{i1} &= \beta_{01} + \beta_{11} x_{1i1} + \ldots + \beta_{k1} x_{ki1} + \varepsilon_{i1}; \\ &\vdots \\ y_{iG} &= \beta_{0G} + \beta_{1G} x_{1iG} + \ldots + \beta_{kG} x_{kiG} + \varepsilon_{iG}. \end{aligned} \quad (6)$$

In a second step, coefficient estimates from the first step become outcome variables in a cluster-level regression with G cases. For example, an analyst interested in the effect of z_g on the slope of x_{1ig} would regress $\hat{\beta}_{11}$ to $\hat{\beta}_{1G}$ on z_1 to z_G.

Separate estimation of the cluster-specific first-step regressions has the important consequence that all β coefficients can vary freely across clusters, without any restrictions on their distribution. Practitioners who want to treat a coefficient as invariant across clusters would have to explicitly introduce such a constraint. That said, two-step approaches are most appropriate for analyzing cross-level interactions. One might be tempted to examine direct context effects by regressing the intercepts of the cluster-specific regressions on contextual variables. But these intercepts are simply predicted values of the outcome variable for the case that all lower-level predictors take the value 0. Because all coefficients are allowed to vary freely across clusters, cluster differences in these predicted values (and hence their relationship with cluster-level predictors) can vary widely depending on which configuration of the covariates this case represents. Another potential issue with the two-step approach is low precision due to the large number of parameters that have to be estimated. This drawback hits hardest when precise estimation of the first-stage β's is not possible due to small cluster sizes or little within-cluster variation.

MULTILEVEL MODELING PRACTICE IN LEADING SOCIOLOGY JOURNALS

Given the choice among these three approaches, how do sociologists analyze multilevel data? To obtain a picture of current practice, we identified and classified all quantitative multilevel analyses that appeared in the *American Journal of Sociology* (AJS), *American Sociological Review* (ASR), and *European Sociological Review* (ESR) in the years 2011 to 2014. Table 1 summarizes key information from this review. The three journals published 541 full-length research articles during the period in question. We classified 117 articles as quantitative multilevel studies.[1] For simplicity, we focus on the 78 articles that used two-level data with lower-level units nested in one type of higher-level unit. Casual inspection suggests our main conclusions also apply to the 39 studies with more complex data structures.

The vast majority of surveyed studies rely on mixed-effects models, with cluster-robust and two-step approaches accounting for only 4 and 3 of the 78 studies, respectively. Analyses of direct context effects and cross-level interactions are equally common in applied

Table 1. Current Multilevel Modeling Practice in Leading Sociology Journals

	Prevalence of Quantitative Multilevel Analyses		Key Features of Quantitative Multilevel Analyses with Hierarchical Two-Level Structure														
			Primary Estimation Approach				Type of Contextual Effect		Proportion of Varying Slopes among Lower-Level Predictors[a]			Number of Clusters			Mean Cases per Cluster		
	Total Number of Articles	Multilevel Analyses with Hierarchical Two-Level Structure	Mixed-Effects Model	Cluster-Robust SE	Two-Step	Mean Number of Lower-Level Predictors[b]	Direct Context Effect Only	Cross-Level Interaction	0	>0 to 50%	> 50%	≤20	21 to 50	>50	≤50	51 to 500	>500
AJS	144	2	2	0	0	15	1	1	1	1	0	0	0	2	1	1	0
ASR	144	20	16	3	1	13.9	9	11	12	7	0	5	5	10	5	9	6
ESR	253	56	53	1	2	13.23	29	27	39	13	2	13	33	10	9	10	37
Total	541	78	71	4	3	13.46	39	39	52	21	2	18	38	22	15	20	43

[a]Without two-step models.
[b]In most comprehensive model.

research. Half the articles examine cross-level interactions in one way or another (some of these studies also investigate direct context effects). Table 1 shows that multilevel analyses differ enormously in terms of the numbers of cases at the higher and lower levels. Among the 78 studies, 18 use fewer than 20 cluster units, and 38 use between 20 and 50. Turning to cluster size, that is, the average number of lower-level units per cluster, roughly 45 percent of the studies have small (50 or fewer) to moderate (50 to 500) cluster sizes. The remaining studies rely on large clusters that contain more than 500 lower-level units, on average. Further analysis reveals that number of clusters and average cluster size are systematically related. Country comparisons, which account for a large portion of multilevel studies, are typically based on few clusters with many lower-level observations per cluster. Applications that can draw on 50 or more clusters mostly study other contextual units, such as neighborhoods or cities. These cases tend to have fewer lower-level observations per cluster.

The typical multilevel model features a substantial number of lower-level predictors. According to Table 1, the average study included 13.46 lower-level independent variables in the most comprehensive model.[2] With mixed-effects models and cluster-robust OLS, the default is to assume that the coefficients of these controls are invariant across clusters. Analysts must explicitly allow them to vary, but our review indicates they hardly ever do. Most studies (52 of the 75 studies using mixed-effects or cluster-robust OLS estimation) assume the effects of all lower-level variables to be invariant across clusters. Among the 23 remaining studies, 21 let the coefficients of at least one, but less than 50 percent, of all lower-level predictors vary across clusters. Usually these predictors are of substantive interest and part of a cross-level interaction. Only 2 of the 75 studies let more than 50 percent of lower-level effects vary. Given this nearly universal practice, the core question of this article is: *what are the consequences of ignoring cross-cluster variation in the coefficients of control variables, and are different multilevel estimation approaches similarly affected?*

IMPLICATIONS OF IGNORING CROSS-CLUSTER DIFFERENCES IN THE EFFECTS OF CONTROLS

We now argue semi-formally that, even under benign conditions, neglect of cross-cluster differences in the effects of lower-level controls reduces the precision of estimated context effects, resulting in unnecessarily wide confidence intervals and low statistical power. To see this, consider the case of two-step estimation. Anticipating one of our empirical illustrations (see the Flexible Multilevel Modeling in Practice section), one might run a separate regression of xenophobia on level of education (and controls) for each of the 28 ESS countries and then regress the estimated effects of education on a country-level variable such as the Human Development Index. More generally, one would first fit G cluster-specific regressions to obtain cluster-specific estimates of an effect of interest, say β_{1g} (the effect of x_{1ig}), and then relate them to a cluster-level, contextual variable, say z_g, to estimate a cross-level interaction between x_{1ig} and z_g.

As is well known (cf. Wooldridge 2014:76ff.), if we omit a confounder x_{2ig} from the regression for cluster g (e.g., we might fail to account for age differences when estimating the relationship between xenophobia and level of education), the expectation of the coefficient estimate on x_{1ig}, say $\tilde{\beta}_{1g}$, will not be the desired β_{1g}, but

$$\mathrm{E}(\tilde{\beta}_{1g}) = \beta_{1g} + \beta_{2g}\frac{\sigma(x_{1ig}, x_{2ig})}{\sigma^2(x_{1ig})}, \quad (7)$$

where $\sigma(x_{1ig}, x_{2ig})$ denotes the covariance of x_{1ig} and x_{2ig} and $\sigma^2(x_{1ig})$ the variance of x_{1ig}. Thus, the more x_{1ig} and x_{2ig} are correlated, and the stronger the effect of x_{2ig} on y_{ig} (i.e., β_{2g}), the more $\mathrm{E}(\tilde{\beta}_{1g})$ will differ from β_{1g}.

The case of interest here, however, is not one where x_{2ig} is omitted completely, but one

where its slope β_{2g} is specified as invariant across clusters, effectively constraining it to its (weighted) average effect $\beta_{2\bullet}$. For example, we might account for age differences among educational groups, but fail to acknowledge that these differences play out very differently in Sweden than they do in Romania. In this scenario, the expectation of the estimated within-cluster slopes of x_1, say β^*_{1g}, will again not be the desired β_{1g}, but[3]

$$\mathrm{E}(\beta^*_{1g}) = \beta_{1g} + (\beta_{2g} - \beta_{2\bullet})\frac{\sigma(x_{1ig}, x_{2ig})}{\sigma^2(x_{1ig})}. \quad (8)$$

Equation 8 shows how erroneously assuming the effect of a lower-level confounder to be invariant will bias cluster-specific estimates of the effects of other variables, unless ($\beta_{2g} - \beta_{2\bullet}$) or $\sigma(x_{1ig}, x_{2ig})$ equals zero. In other words, estimated education-related differentials in xenophobia for country g will be biased, unless education and age are uncorrelated in g or the effect of education on xenophobia in g equals the average effect across the 28 European countries. As for the former condition, it is well known that level of education and age are correlated in advanced economies (due to cohort differences in educational attainment). As for the latter, our illustrative analysis (see the Flexible Multilevel Modeling in Practice section) demonstrates that it is unlikely to hold. We find that the relationship between xenophobia and age varies considerably across the 28 countries in our sample. More importantly, this also holds for most other combinations of outcome measures and predictors we consider.

If heterogeneity in the slopes of controls is purely random (i.e., unsystematic with respect to the contextual variable of interest), improper adjustment for lower-level controls should not introduce bias into estimated context effects; but it does add noise to the cluster-specific estimates of the lower-level relationship of interest. Thus, our argument suggests that erroneously assuming the effects of lower-level controls to be invariant reduces the precision of estimated context effects. Moreover, the loss of precision should increase with the *number of lower-level controls whose effects vary* and with the *extent of cross-cluster variation in their coefficients*.

Asymptotically (i.e., as the number of clusters approaches infinity), the additional noise introduced by the invariant coefficients assumption is no major concern—intuitively this is because the cluster-specific biases cancel each other out in large samples. But in small samples it can substantially increase the risk that a single study will produce very misleading estimates of the contextual effect of interest. Accordingly, the *number of cluster units* is an influential factor.

So far, we have focused on the *cross-level interaction* case. We argued that cluster-specific estimates of β_{1g} will be biased when one erroneously constrains the slope of a control variable to its average effect. How about the *direct effect of a contextual factor* on y_{ig}? It is important to acknowledge that the inclusion of lower-level controls serves different purposes in the two cases. In the cross-level interaction case, the primary goal is to purge the cluster-specific relationships between x_{1ig} and y_{ig} of differences that are due to confounding variables. In the case of a direct context effect, the goal is to adjust cluster differences in the level of y_{ig} for variability that reflects compositional differences with respect to the lower-level controls. This adjustment will be imperfect if the model erroneously assumes the slopes of lower-level controls to be invariant across clusters. Thus, the precision costs associated with erroneously assuming invariant coefficients should increase with the extent of *compositional differences across clusters*, particularly in the case of direct context effects.

How can the earlier-discussed multilevel modeling approaches tackle heterogeneity in the slopes of controls? In the cluster-robust OLS setting, this is difficult and precision losses seem largely inevitable. Mixed-effects models can accommodate heterogeneous coefficients of controls by specifying random slopes on control variables; but our review of current practice shows this is rare in practice.

Two-step approaches automatically accommodate any number, strength, and structure of cluster-varying control slopes during the first step. Yet, it is not clear if this advantage is large enough to outweigh the costs of estimating many more parameters than mixed-effects or cluster-robust OLS estimation.

MONTE CARLO SIMULATION STUDY

We conduct Monte Carlo simulations to explore the implications of neglecting cluster heterogeneity in the coefficients of control variables for the estimation of context effects. Monte Carlo simulations are a powerful tool for assessing different estimation approaches under controlled and ideal conditions, when purely analytic comparisons are not feasible. The basic idea is to repeatedly simulate data with a known data-generating process (DGP) and to apply alternative estimation approaches to each simulated dataset (also referred to as replications). The performance of the different approaches over a large number of datasets allows one to compare their statistical properties. By repeating this exercise for different DGPs, one can better understand how key features of actual applications (e.g., the number of clusters) shape the relative performance of the alternative estimators.

Data-Generating Processes

Equation 9 describes the basic data-generating process for our simulations. As before, y_{ig} is the outcome for individual i in cluster g, which is determined by an intercept β_{0g}, six lower-level predictors[4] x_{1ig} to x_{6ig} with associated coefficients β_{1g} to β_{6g}, and an error term ε_{ig}:

$$y_{ig} = \beta_{0g} + \beta_{1g}x_{1ig} + \beta_{2g}x_{2ig} + \beta_{3g}x_{3ig} + \beta_{4g}x_{4ig} + \beta_{5g}x_{5ig} + \beta_{6g}x_{6ig} + \varepsilon_{ig}. \quad (9)$$

The g indices on the intercept and slope terms in Equation 9 indicate that their size may vary across clusters. We are interested in how well different approaches estimate the coefficients of cluster-level (i.e., contextual) variables when such variation is present in the data. We study two broad variants of the DGP in Equation 9, which resemble the two types of research questions addressed in applied research. Both variants include one contextual factor z_g. In the first variant, which we term DGP-DCE (with DCE for direct context effect), z_g affects only the intercept in Equation 9; formally

$$\beta_{0g} = \gamma_{00} + \gamma_{01}z_g + \nu_{0g}. \quad (10)$$

The second variant of the DGP—the DGP-CLI—has a cross-level interaction: z_g also affects the slope of x_{1ig}; formally

$$\beta_{1g} = \gamma_{10} + \gamma_{11}z_g + \nu_{1g}. \quad (11)$$

In both DGPs we set all fixed effects to 1 (i.e., all γ's, including the effects of interest, γ_{01} and γ_{11}). Our main research question is how well different approaches estimate γ_{01} and γ_{11} when the coefficients of lower-level control variables are not constant across clusters. For both DGP-DCE and DGP-CLI, we consider several variants. Adopting a common term from the Monte Carlo literature, we refer to the different variants as experimental conditions. The different experimental conditions follow directly from our review of published research and from the semi-formal argument presented in the previous section.

A first factor we vary is the number of lower-level control variables (i.e., x_{2ig} to x_{6ig}) whose effects differ across clusters. Specifically, we let the coefficients of these variables depend on random effects ν_{2g} to ν_{6g}: $\beta_{kg} = \gamma_{k0} + \nu_{kg}$ for $k \in \{2,3,4,5,6\}$. We generally start with a baseline set-up that includes all lower-level controls but has none of their coefficients vary across clusters. We then introduce random cross-cluster variation by setting the standard deviations of the ν_{kg}'s to a value greater than zero for one, three, and finally all five lower-level controls. We consider two values for the standard deviation of the ν_{kg}'s: .2 (20 percent of the average effect) and 1 (100 percent of the average effect). These values fall into the lower and upper

Table 2. Monte-Carlo Simulations: Dimensions of Comparison

Dimension	Implementation	Levels
Cluster Heterogeneity I	Number of random effects on control variables	0 1 3 5
Cluster Heterogeneity II	Standard deviation of random effects	.2 1
Type of Application/Setting	Number of clusters G and individual units per cluster N_g	15 countries: $G = 15$; $600 \leq N_g \leq 2000$ 25 countries: $G = 25$; $600 \leq N_g \leq 2000$ 50 cities: $G = 15$; $70 \leq N_g \leq 130$
Compositional Differences among Clusters	Standard deviation of cluster means for control variables	No compositional differences among clusters 15 percent of variance in control variables between clusters 50 percent of variance in control variables between clusters

ends of the range observed in our illustrative empirical analyses (see the Flexible Multilevel Modeling in Practice section). The random effects related to the contextual predictor of interest z_g (i.e., v_{0g} in DGP-DCE and v_{0g} and v_{1g} in DGP-CLI) always have standard deviations of .6. The random effects are multivariate normal with a random correlation matrix. We provide formal representations of the full DGPs in Part C of the online supplement.[5]

For both DGP-DCE and DGP-CLI, we further manipulate the number and size of clusters. We consider three scenarios that resemble those typically found in applied work. The first two mimic country-comparative applications with few clusters (15 and 25) and many observations (uniformly distributed between 600 and 2,000) per cluster. The third setting features more clusters (50) but fewer observations per cluster (uniformly distributed between 70 and 130). These numbers are typical of studies analyzing clusters such as cities or census tracts. The last dimension we modify is the extent of cross-cluster compositional differences with respect to lower-level variables. We consider three levels where 0 percent (no compositional differences), 15 percent, and 50 percent of the total variance in lower-level predictors is between clusters. As with the random slopes on controls, compositional differences are random and unrelated to the contextual variable z_g.[6]

Table 2 lists the four dimensions of the DGPs that we vary, along with their different levels. There are 72 (= 4 × 2 × 3 × 3) experimental conditions for each of the two broad variants of the DGP (DGP-DCE and DGP-CLI). We ran simulations for all experimental conditions, obtaining 10,000 replications each to minimize Monte Carlo error. We focus on the most important patterns here and provide additional results in the online supplement. Part D of the online supplement describes how we implemented the Monte Carlo simulations in the software package R. Code for replicating the simulations is available with the online supplements.

Estimators Compared

We consider four estimators. The first is a mixed-effects multilevel model that always assumes the coefficients of lower-level control variables to be invariant across clusters (ME-Invariant). As noted in our review of multilevel modeling practice, this currently is by far the most common estimator in applied sociological research. When applied to DGP-DCE, this model is a random-intercept model. When applied to DGP-CLI, it is a model with

a random intercept and one random slope, namely on the predictor of interest x_{1ig}. The second estimator is a mixed-effects model with the correct random-effects structure given the version of the DGP it is applied to. Thus, it is identical to ME-Invariant in the baseline setting, but contains one, three, and five additional random effects in the more complex variants of the DGPs. We impose no constraints on the correlations between the random effects, which are estimated from the data. We refer to this model as ME-Correct. The third estimator is pooled OLS with cluster-robust standard errors (OLS-Cluster), which like ME-Invariant does not allow for cluster differences in the coefficients of controls. The last estimator (two-step-FGLS) is a version of the two-step estimator that implements the Feasible Least Squares approach described by Lewis and Linzer (2005).[7] We consider two-step-FGLS only when estimating the cross-level interaction between z_g and x_{1ig} (i.e., γ_{11} in DGP-CLI).

Estimation of mixed-effects models can run into convergence problems, that is, the optimizer may fail to identify the maximum of the likelihood function. Fortunately, separate analysis of replications with and without convergence warnings does not suggest that our main conclusions are sensitive to the convergence status of the mixed-effects estimators. We therefore present simulation results with convergent and non-convergent solutions pooled together. Part E, Section E.5, of the online supplement provides disaggregated results by whether convergence warnings occurred or not.

Quantities of Interest

A first quantity of interest in most simulation studies is bias in point estimates. An estimator is unbiased if parameter estimates are equal to their true values in expectation. In the present study, all estimators produce unbiased estimates of context effects because we consider only unsystematic differences in the effects of controls. Here, we simply note this fact and do not discuss bias further (results are available upon request).

Our main interest lies with the precision of point estimates. Intuitively, precision refers to the variation of point estimates around the corresponding true values over repeated samples. That is, it measures how reliably an estimator approximates the true parameter values. Other things—in particular, parameter bias—being equal, one would prefer a more precise estimator to a less precise one, because the former tends to provide more accurate estimates of the true effects. A standard measure for characterizing the precision of an estimator is the Root Mean Squared Error (RMSE), with smaller values implying greater precision. Let $\hat{\gamma}_m$ denote the point estimate for the mth simulated dataset (or replication) and let γ denote its true value (i.e., its value in the DGP). With a total of M Monte Carlo replications, the RMSE equals

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(\hat{\gamma}_m - \gamma\right)^2}.$$

Another important question is whether the different approaches correctly estimate statistical uncertainty. We therefore investigate whether the actual coverage rate of two-sided 95 percent confidence intervals equals their nominal coverage rate. In other words, we examine if these intervals cover the true parameter value in more or less than 95 percent of the 10,000 Monte Carlo replications. Let $C_{95}(m)$ equal one if the 95 percent confidence interval for the mth replication includes the true value of the parameter of interest and zero otherwise. Then coverage is defined as

$$\text{Coverage} = \frac{1}{M}\sum_{m=1}^{M} C_{95}(m).$$

If coverage is greater than 95 percent, confidence intervals are too large and overconservative. Hypothesis tests will retain the null hypothesis of no effect too often. By contrast, if coverage is below 95 percent, confidence intervals are too narrow and null hypotheses rejected too frequently. We use a t-distribution with $G - 2$ degrees of freedom to identify the limits of analytic 95 percent confidence intervals for the coefficient of the

contextual predictor z_g and its cross-level interaction with x_{1ig} (where G represents the number of clusters and 2 is subtracted to account for the degrees of freedom consumed by the intercept and the cluster-level predictor; see Elff et al. 2016; Raudenbush and Bryk 2002:280).

Simulation Results I: Consequences of Heterogeneity in the Effects of Controls for Estimates of Direct Context Effects

We begin with the estimation of direct context effects. Figure 1 displays precision in terms of the RMSE, expressed in percent of the true effect. Results are for 25 clusters with 600 to 2,000 lower-level observations each, numbers that are typical of country-comparative studies. The figure consists of six panels. Within each panel, we change the number of variables with varying coefficients from left to right. The three lines thus show how the precision of the different estimators evolves as the DGP incorporates increasing heterogeneity in the effects of the five control variables. The coefficients of lower-level controls have cross-cluster standard deviations of .2 (20 percent of the average effect) in the top row and 1 (100 percent of average effect) in the bottom row. For better readability, the scale is larger in the top than in the bottom row. Across the different columns, we vary the extent of between-cluster compositional differences with respect to the lower-level controls. Note that we use a logarithmic scale on the *y*-axis (RMSE). The logarithmic scaling means that a given distance on the scale always corresponds to the same *relative* change in RMSE (i.e., the distance between 10 and 20 is identical to that between 20 and 40). The value labels attached to the axis are directly interpretable without further transformation.

A consistent result in Figure 1 is that ignoring cross-cluster differences in the coefficients of lower-level variables can induce substantial and unnecessary uncertainty into estimates of direct contextual effects. Within each panel, the precision gap (i.e., the difference in RMSE) between ME-Correct and the other two estimators widens as we manipulate the DGP to include random slopes on one, three, and eventually all five lower-level control variables. Focusing on a concrete example, panel 1a (i.e., the case of no compositional differences and a cross-cluster standard deviation of .2 for controls with varying effects) shows that a mixed-effects model that correctly specifies random slopes on five control variables with heterogeneous effects (ME-Correct) has an RMSE of 11.2 percent. The models that erroneously assume invariant coefficients, ME-Invariant and OLS-Cluster, have RMSEs of 12.9 and 13.4 percent, respectively. More generally, the RMSE of ME-Correct tends to be smaller by a factor of .7 to .9 in the less extreme experimental conditions, that is, ME-Correct is 10 to 30 percent more precise than ME-Invariant.

To see that these differences are substantial, note that the RMSEs are essentially Monte Carlo approximations to the true standard errors of the estimators and that increasing the sample size by a factor of a reduces the standard error by a factor of approximately $\sqrt{a}$. Thus, even using the lower-bound estimate of a 10 percent reduction in the RMSE, ME-Invariant would need roughly 31 rather than 25 cluster units to produce equally precise estimates of the context effect as ME-Correct (i.e., the sample would have to be larger by a factor of $1.23 = 1/.9^2$). Comparativist scholars would certainly applaud an addition of six countries to a cross-national study such as the European Social Survey and see it as a considerable increase in terms of statistical power for multilevel analysis. Our Monte Carlo results suggest that a comparable increase can be achieved by adequately accounting for cross-cluster differences in the effects of lower-level variables.

In many of the other experimental conditions, the precision gains from using ME-Correct rather than OLS-Cluster or ME-Invariant are even larger. The most extreme case in Figure 1 occurs in panel 3b, where ME-Correct has an RMSE of 13.5 percent, and the corresponding values for ME-Invariant and OLS-Cluster are roughly 3.5 times as large at 48.7 percent and 51.7 percent of the true effect.

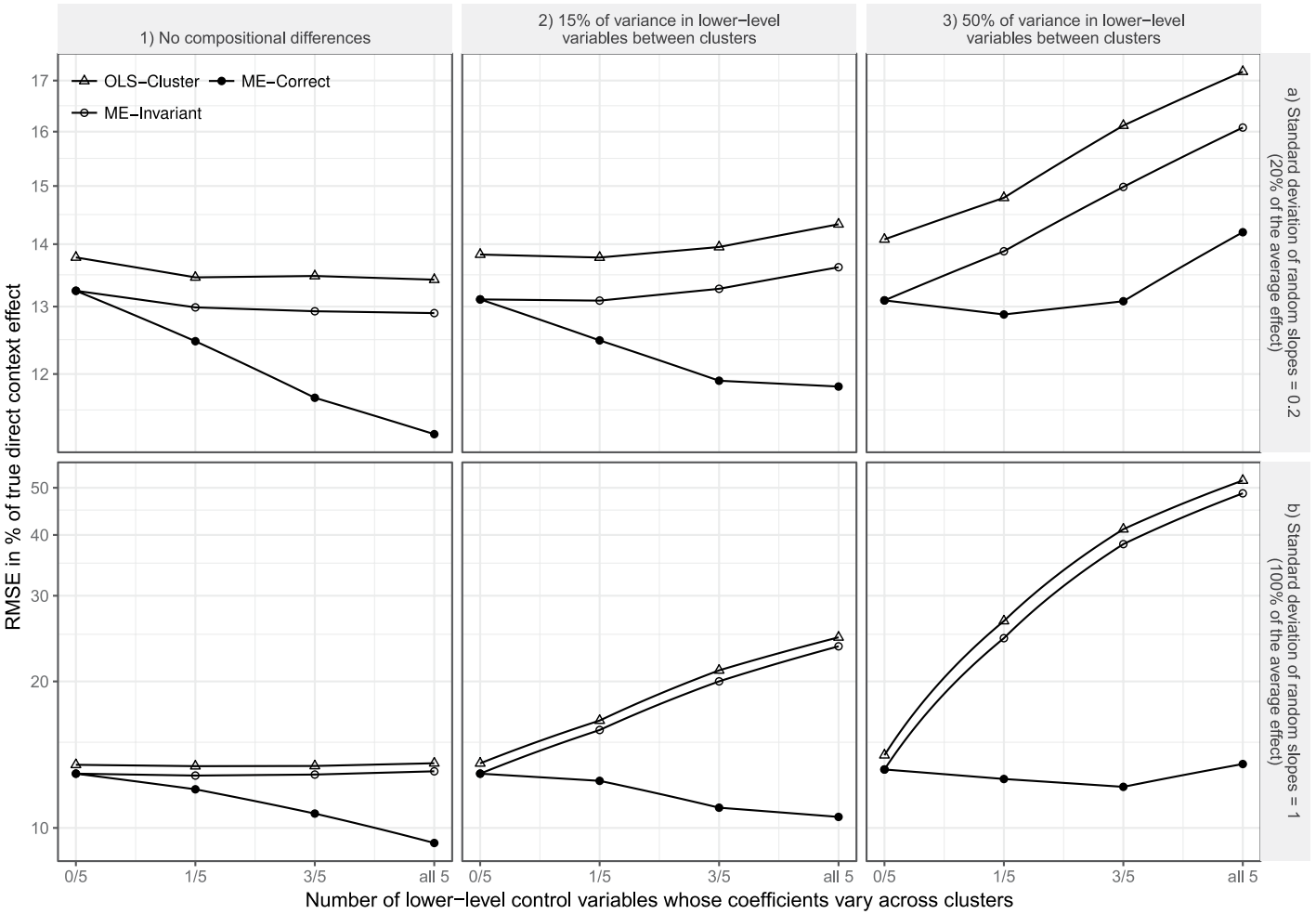

Figure 1. Precision of Estimated Direct Context Effects by Cross-Cluster Compositional Differences and Variation in the Coefficients of Lower-Level Controls

Note: The *y*-axis has a logarithmic scale, so a given distance on the scale always corresponds to the same relative change in RMSE. The value labels attached to the axis are directly interpretable without further transformation.

Overall, OLS-Cluster and ME-Invariant perform quite similarly in Figure 1. ME-Invariant is generally slightly more accurate than OLS-Cluster, but this difference is negligible compared to the advantages of ME-Correct that emerge in many settings. This suggests that whether a model assumes invariant coefficients of lower-level variables will often have a much larger impact on precision than will the method of estimation used (mixed-effects versus pooled OLS).

Detailed comparisons across the six panels in Figure 1 reveal how key aspects of the DGP affect the magnitude of precision losses. Comparing the top with the bottom row shows that more pronounced heterogeneity in the effects of controls exacerbates the costs of making the invariant coefficients assumption. For any given number of control variables with varying coefficients, the precision gap between ME-Correct and the other two estimators is larger when the varying coefficients have a standard deviation of 1 (bottom row) rather than .2 (top row). Another crucial pattern in Figure 1 concerns the extent of compositional differences among clusters. In the (unrealistic) case that there are no compositional differences (left column), this gap is very small throughout and primarily due to ME-Correct becoming more precise as the number of random slopes increases. Apparently, ME-Correct can exploit information about systematic correlations among the random slopes to arrive at somewhat better estimates of the parameter of interest.[8] Greater compositional differences clearly amplify the precision gap between ME-Correct and the other two estimators. The effect of growing compositional differences on the precision gap is dramatic when we set the standard deviations of the varying coefficients on lower-level controls to 1 in the DGP (bottom row). It is smaller, but still sizable, when we assume limited cross-cluster variation (standard deviation of .2, top row).

Figure 2 explores how the number of higher- and lower-level units affects the relative performance of the estimators. For simplicity, we now consider only the case of moderate compositional differences, with 15 percent of the variance in lower-level variables being among clusters. Figure 2 indicates that the invariant coefficients assumption is costlier with few clusters (here 15 or 25 countries). With 50 cities, the random errors introduced by erroneously assuming invariant coefficients tend to average out. Only in the case of strong variability in the true effects of lower-level controls (panel 3b) do we still find ME-Invariant and OLS-Cluster to be substantially less accurate than ME-Correct.

We now turn to statistical inference. Figure 3 shows results for the case of moderate compositional differences (15 percent of variance in lower-level predictors among clusters). Consistent with previous research (Cameron and Miller 2015), we find that OLS-Cluster underestimates statistical uncertainty when the number of clusters is small. The problem becomes less severe as the number of clusters increases, but undercoverage remains noticeable even in the 50 cities condition. Further analysis shows that the extent of undercoverage is positively related to the extent of compositional differences (see Figure E.3.1 in the online supplement). In combination with its relatively poor performance in terms of accuracy, these results suggest that OLS with conventional cluster-robust standard errors would not be a good choice in most applications.

ME-Invariant consistently yields valid inference—even in the extreme case of only 15 clusters and five (unspecified) varying coefficients in the DGP. This result seemingly contradicts Stegmueller (2013) and Bryan and Jenkins (2016), who conclude that linear mixed-effects multilevel models require at least 20 to 25 clusters for unbiased inference on the coefficients of context variables. The reason for this divergence is that we use restricted maximum likelihood estimation and a *t*-distribution with the approximately correct $G - 2$ degrees of freedom rather than the normal distribution to identify confidence limits (Elff et al. 2016).

Results for ME-Correct are sobering in that it generally produces confidence intervals that are too narrow. Thus, there seems to be a trade-off between precision and valid statistical inference.

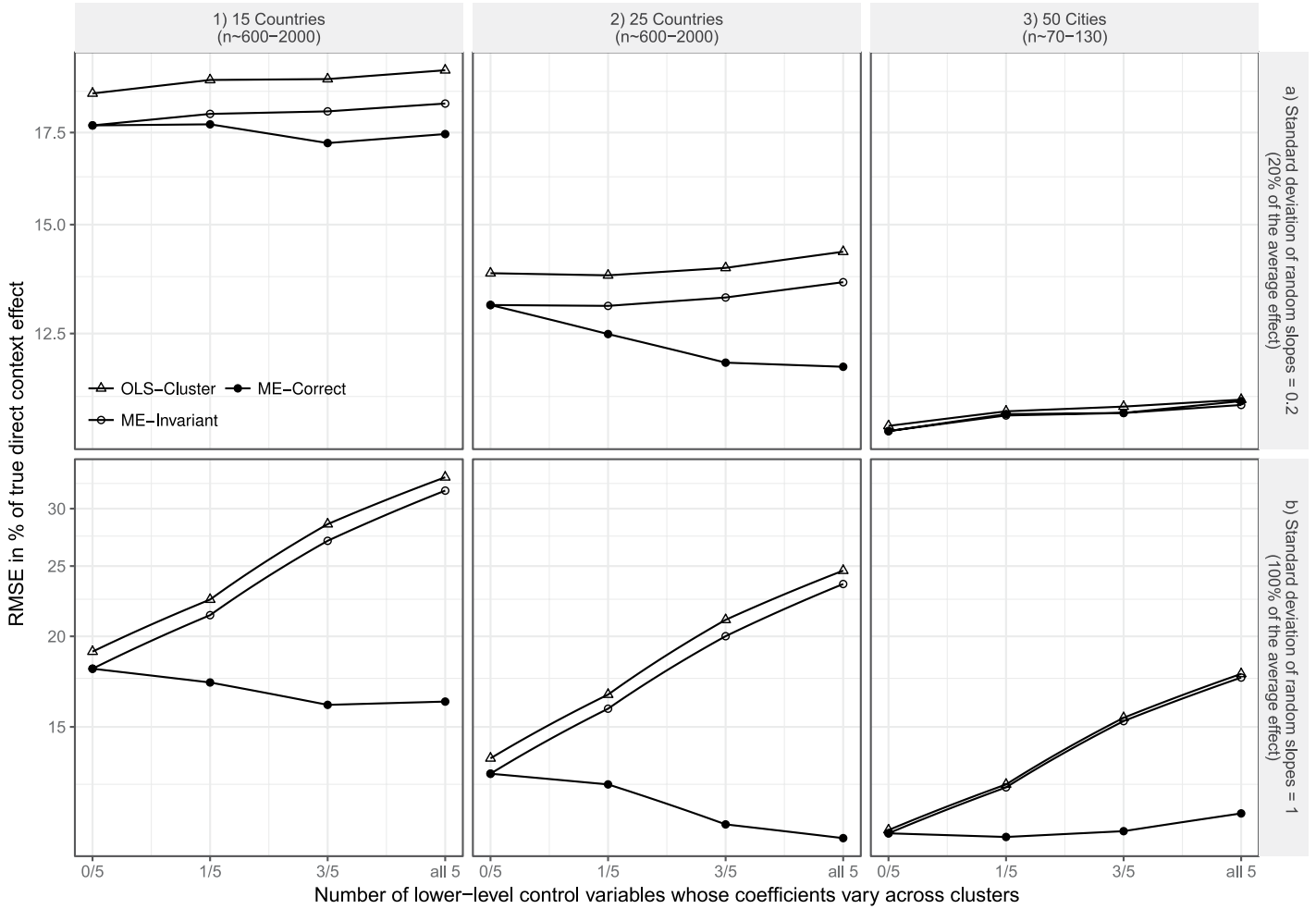

Figure 2. Precision of Estimated Direct Context Effect by Number and Size of Clusters and Cross-Cluster Variation in the Coefficients of Lower-Level Controls

Note: The *y*-axis has a logarithmic scale, so a given distance on the scale always corresponds to the same relative change in RMSE. The value labels attached to the axis are directly interpretable without further transformation.

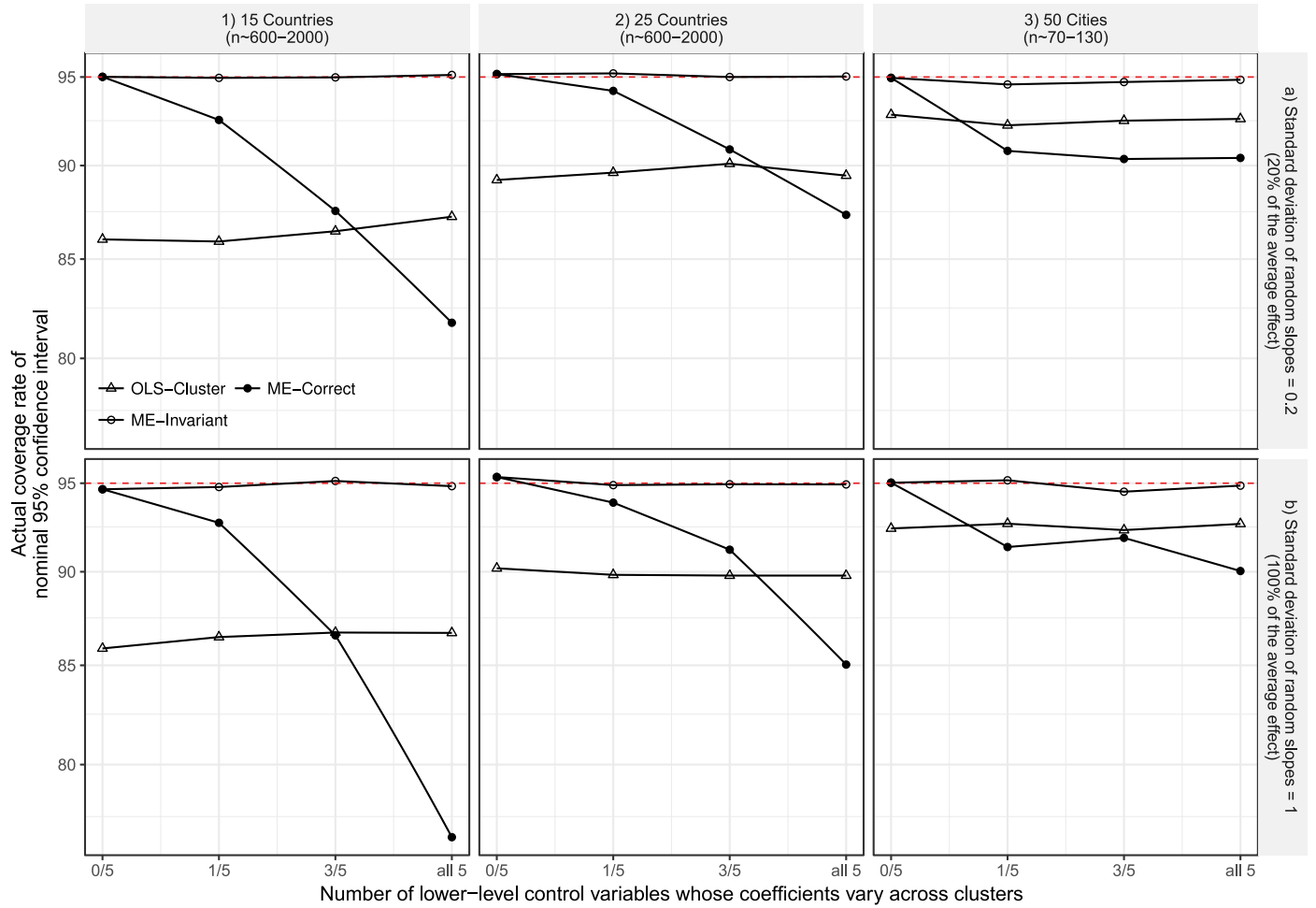

Figure 3. Statistical Inference for Direct Context Effects by Number and Size of Clusters and Cross-Cluster Variation in the Coefficients of Lower-Level Controls

Undercoverage worsens as the number of correctly specified random slopes increases. Consider the case of 15 countries and a standard deviation of .2 for the random slopes on controls (panel 1a in Figure 3). In this setting, the actual coverage rate of nominal 95 percent confidence intervals is approximately 92.5 percent when only one of the lower-level controls has varying coefficients. When the slopes of all five controls vary across clusters (and ME-Correct specifies five random slopes), coverage drops to approximately 81.8 percent. Only the findings for the 50 cities scenario (panels 3a and 3b) deviate from the pattern that undercoverage increases with the number of random slopes, but as we further explore in the online supplement (Section E.5), this apparent irregularity is driven by replications where ME-Correct ran into convergence problems. Figure 3 further indicates that the undercoverage problem attenuates as the number of clusters increases (columns). The extent of cross-cluster variation in the slopes of lower-level variables seems to matter very little (top versus bottom row).[9] Finally, additional analysis reveals that stronger compositional differences exacerbate the problem (see Figure E.3.1 in the online supplement).

A likely explanation for the inferential deficiencies of ME-Correct is that its complex random-effects structure may sometimes be overspecified: when the effects of all five lower-level controls vary, the covariance matrix has 21 elements (the variances of the random intercept and five random slopes as well as 15 covariances). Overspecification issues should increase with the number of random slopes and decrease with the number of clusters, which is exactly what Figure 3 shows. This raises questions of model simplification and model selection that we will address in the context of the empirical illustrations (see the Flexible Multilevel Modeling in Practice section).

Another important question is whether valid inference for complex mixed-effects models is possible even in the presence of overspecification. In the final step of the Monte Carlo analysis, we present simulation results suggesting that non-parametric bootstrapping may do the trick.

Simulation Results II: Consequences of Heterogeneity in the Effects of Controls for Estimates of Cross-Level Interactions

Now we turn to cross-level interactions where a contextual factor moderates the association between a lower-level predictor and an outcome variable (i.e., we focus on estimation of γ_{11} in DGP-CLI; see Equation 11). Here we additionally consider two-step-FGLS, which avoids the invariant coefficients assumption by design. For brevity, we report results for different levels of compositional differences in Figure E.2.1 in the online supplement. The main finding is that even without compositional differences among clusters, erroneously assuming invariant control slopes can severely reduce the precision of estimated cross-level interactions. Figure 4 illustrates how precision depends on the extent of variation in the slopes of lower-level controls (top versus bottom row) and the number of clusters and lower-level units per cluster (columns). We show results for moderate compositional differences (15 percent of variance in lower-level variables between clusters).

In one important respect, Figure 4 conveys the same messages as Figures 1 and 2. Ignoring cross-cluster heterogeneity in the coefficients of lower-level variables can result in very imprecise estimates. Except in the baseline case where all controls have the same effect in all clusters, OLS-Cluster and ME-Invariant have larger RMSE than does ME-Correct. When variability in the slopes of controls is high (bottom row), they also perform worse than two-step-FGLS in the country scenarios. The precision gap between OLS-Cluster and ME-Invariant and the other estimators primarily stems from the fact that the former become less accurate as the number of controls with varying coefficients increases. As in the case of direct context effects, we also find that the costs of assuming invariant coefficients are largest when the number of clusters is small.

Overall, precision losses are quite similar to the direct context effects case (see Figure 2). As an example, consider the case with varying coefficients on all five controls in

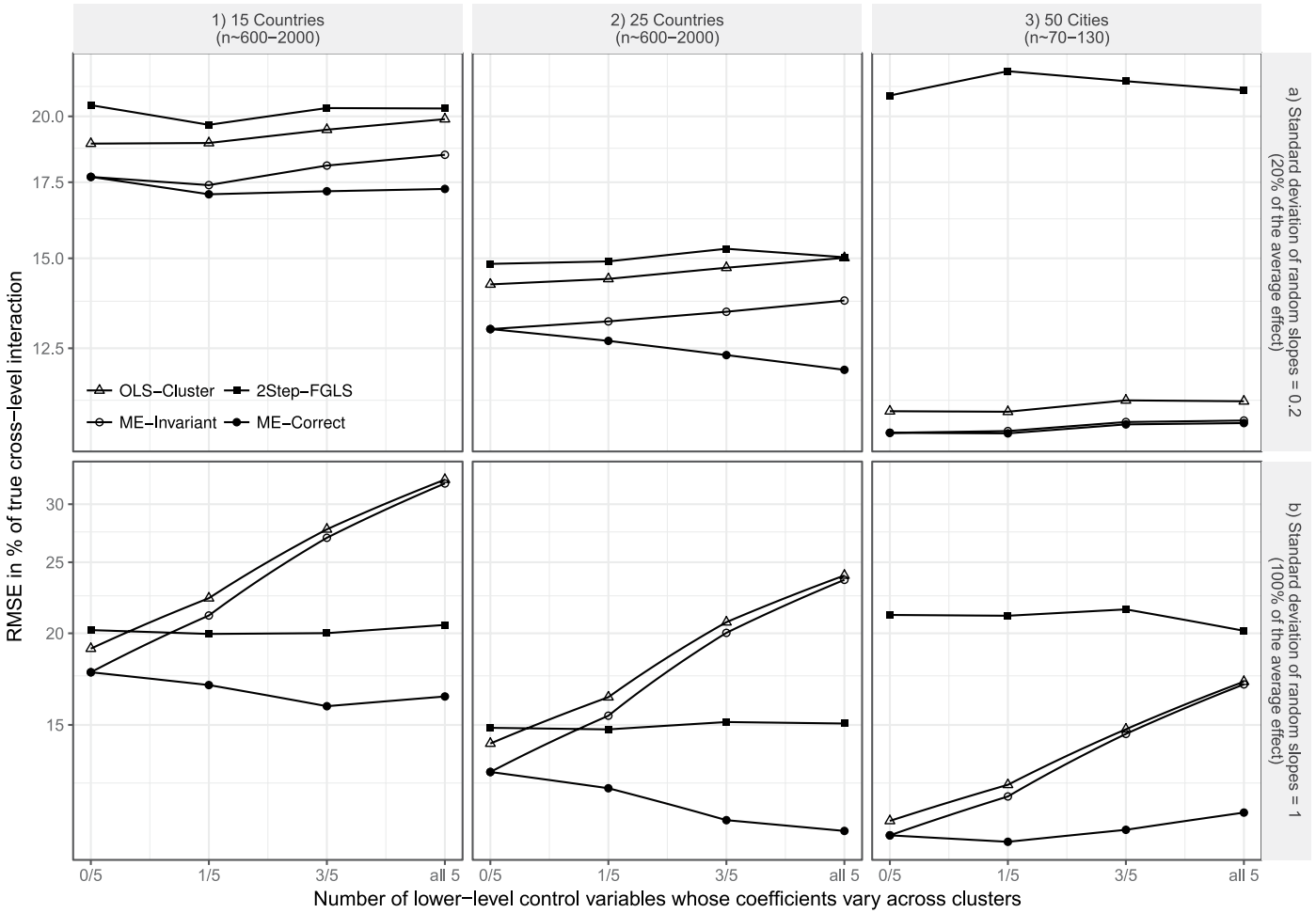

Figure 4. Precision of Estimated Cross-Level Interaction Effects by Number and Size of Clusters and Cross-Cluster Variation in the Coefficients of Lower-Level Controls

Note: The *y*-axis has a logarithmic scale, so a given distance on the scale always corresponds to the same relative change in RMSE. The value labels attached to the axis are directly interpretable without further transformation.

panel 1b of Figure 4 (i.e., the case of 15 countries and a cross-cluster standard deviation of 1 for the effects of controls). Here the RMSE is 16.4 percent for ME-Correct and 32.0 percent for ME-Invariant. To make things more concrete, we can use the fact that the RMSE is effectively the standard deviation of point estimates around the true parameter value. This implies that in 20 percent of all potential applications, ME-Correct will miss the true value of the coefficient on the cross-level interaction by *at least* 21.0 percent.[10] The corresponding number for ME-Invariant is 41.0 percent. In 10 percent of all applications, the two will miss the true value by at least 27.0 percent and 52.6 percent, respectively.

How does two-step-FGLS perform? By design it neither loses nor gains precision as the number of control variables with varying effects increases. In all panels of Figure 4, we essentially see a straight line with some random fluctuation. The level of the line, however, differs considerably across the panels, both in absolute terms and, more importantly, also relative to the alternative estimators. When there is limited heterogeneity in the effects of lower-level controls (top row), two-step-FGLS generally is the least efficient estimator. The picture is different in the bottom row, where the slopes of lower-level controls with heterogeneous coefficients have a cross-cluster standard deviation of 1 rather than .2. In the country conditions (panels 1b and 2b), two-step-FGLS clearly outperforms ME-Invariant and OLS-Cluster as soon as we include three control variables with heterogeneous coefficients in the DGP. Yet, two-step-FGLS remains less precise than ME-Correct. The latter not only combines a flexible specification of random slopes with a more parsimonious and efficient estimation approach, but it also gains precision because of the information inherent in correlated random slopes. An unambiguous result in Figure 4 is that two-step-FGLS produces very imprecise estimates in the cities setting, even when there is marked cross-cluster heterogeneity in the effects of controls (panel 3b). The costs of fitting a separate first-level regression based on only 70 to 130 observations for each cluster clearly are substantial.

Turning to statistical inference, results for the mixed-effects estimators and OLS-Cluster resemble those for direct context effects (see Figure E.4.1 in the online supplement for visualizations of these results). OLS-Cluster generally produces confidence intervals that are too narrow, as does ME-Correct, particularly when there are several (correctly specified) random slopes in the model. Again, having only a few clusters amplifies these problems. By contrast, both ME-Invariant and two-step-FGLS consistently achieve correct inference.

In summary, our simulation results suggest that neglecting cross-cluster differences in the effects of controls can substantially reduce the precision of estimated context effects. This holds for direct context effects as well as for cross-level interactions. In many experimental conditions, mixed-effects multilevel models with a flexible random-effects structure yield much more precise point estimates than do invariant specifications that assume the effects of lower-level variables to be constant across clusters. For the cross-level interaction case, two-step estimation is an easy-to-implement alternative that allows coefficients to vary freely across clusters. It shows good performance when clusters are sufficiently large, but generally remains less precise than a mixed-effects model with the correct random-effects structure.

More precise estimates reduce the risk that a given analysis severely misrepresents the true effect of a contextual variable. In addition, they will more often lead to rejection of the null hypotheses of no effect when it is in fact wrong (greater statistical power/lower type-II error rates). The benefits of using a flexible specification are largest when heterogeneity is pronounced (many and strongly variable coefficients) and when analysts have only a few clusters at their disposal. In the DCE case, the precision gains also increase with the extent of compositional differences among clusters. These conditions are most typical of country-comparative studies. Unfortunately, analytic confidence intervals for complex mixed-effects models can be severely anticonservative, so model simplification or other methods of inference are needed.

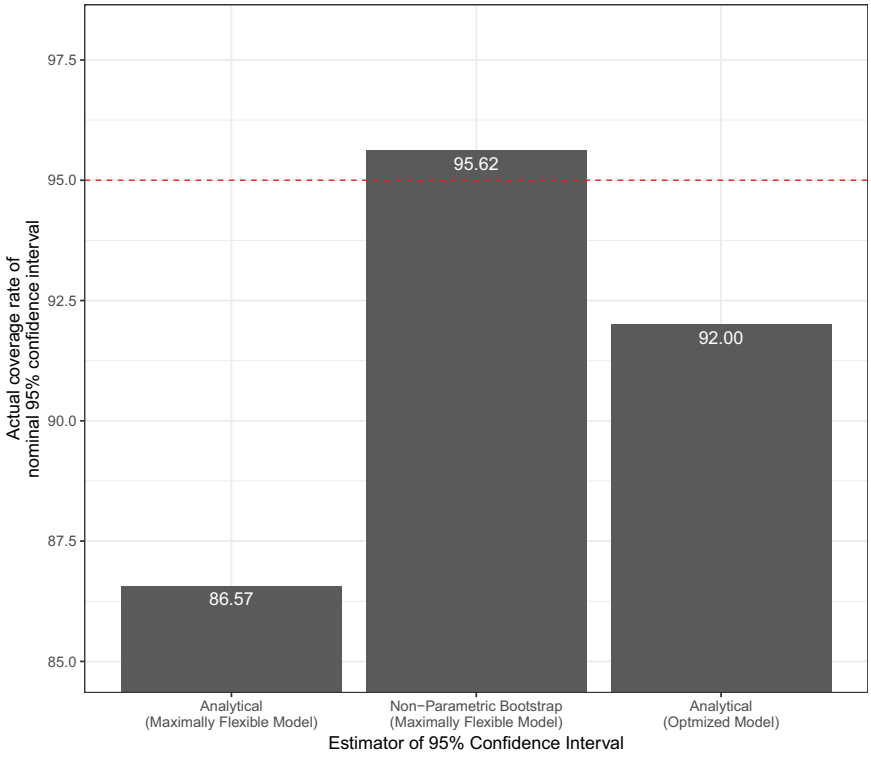

Figure 5. Coverage Rates for Analytic and Bootstrap-Based 95 Percent Confidence Intervals

Bootstrap Inference for Complex Mixed-Effects Models

Given the problems with conventional statistical inference for ME-Correct (see Figure 3), we ran additional simulations to evaluate the performance of non-parametric cluster bootstrap confidence intervals (Goldstein 2011). In Part F of the online supplement, we describe the bootstrap procedure in detail and show how to implement it in R and Stata. Because of the large computational burden, we focus on one experimental condition where analytic confidence intervals suffer from severe undercoverage: the case of 15 countries, moderate compositional differences (15 percent of variance between clusters), and three controls with varying slopes, each with a standard deviation of 1. In this condition, ME-Correct is a model with random slopes on all three controls and no constraints on the correlations among them. We ran 5,000 Monte Carlo replications and 2,000 bootstrap replications per Monte Carlo replication.[11]

The rightmost bar in Figure 5 shows the coverage rate of analytic confidence intervals for an optimized specification. We elaborate on this at the end of the next section. The crucial comparison at this point is between the actual coverage rates of analytic (leftmost bar in Figure 5) and non-parametric cluster bootstrap confidence intervals (middle bar) for ME-Correct. Figure 5 sends a clear message: the non-parametric cluster bootstrap performs well. It improves dramatically on the much too low coverage rate of analytic confidence intervals (85.67 percent; cf. panel 1b in Figure 3). At 95.62 percent, the coverage rate of the bootstrapped confidence interval is close to the nominal level. If anything, it appears to be slightly over-conservative (in fact the difference to the nominal coverage rate of 95.00 percent is barely significant at the 5 percent level). Further research on this

issue is required, but it appears that non-parametric bootstrapping can effectively resolve the inferential limitations of ME-Correct.

FLEXIBLE MULTILEVEL MODELING IN PRACTICE: AN ILLUSTRATIVE ANALYSIS

As a final step, we now explore the transferability of the Monte Carlo results to real-life settings by presenting illustrative analyses of five outcome variables based on the European Social Survey (ESS Round 6 2016), a dataset widely used in cross-national comparative research.

To foreshadow the results, we show that the coefficients of six standard individual-level (control) predictors, such as gender and age, differ substantially across ESS countries. The extent of cross-country variability is broadly consistent with the experimental conditions used in the Monte Carlo simulations. However, practitioners would be mistaken to simply estimate maximally flexible mixed-effects models with random slopes on all lower-level predictors (and no constraints on the correlations among them). Such models tend to yield less precise estimates of context effects than the invariant specification. A likely reason is that maximally flexible models will often suffer from overparameterization. Small cluster-level samples limit the number of random effects (and correlations among them) that can be estimated reliably. In such a situation, random-effects structures need to be selected carefully. Including random slope terms with little actual variance may do more harm than good. Recall that the most complex mixed-effects model investigated in the Monte Carlo simulations (i.e., ME-Correct) did not generally specify random slopes on all controls. It did so only for the predictors whose effects actually varied across clusters, which we happened to know because the data were artificially created. Even so, the problems with analytic inference for ME-Correct suggest it may have stretched the (simulated) data beyond the limits in some situations. The presence of a random effect in the underlying DGP does not mean that a finite dataset necessarily contains enough information to reliably model it.

In real-life applications, one must strike the right balance between flexibility and parsimony to reap the increased precision found in the Monte Carlo analysis. Fortunately, practitioners can rely on the well-established toolbox of regression diagnostics and model selection criteria. Building on these techniques, we find that models whose random-effects structure has been simplified according to the procedure outlined by Bates, Kliegl, and colleagues (2015) tend to produce more precise estimates of context effects than do invariant specifications.

An Illustrative Analysis of the European Social Survey

Taking its cue from a sociological classic, our empirical illustration studies how five outcome variables—generalized trust, xenophobia, occupational status, homophobia, and fear of crime—depend on the Human Development Index (HDI; United Nations Development Programme 2015) as a broad indicator of modernization across a sample of 28 European countries.[12] All outcome variables are standardized to have a mean of zero and a standard deviation of one. In keeping with the two types of research questions identified earlier, we consider both a direct context effect of the HDI and its cross-level interaction with having a high level of education. For instance, we investigate whether overall levels of generalized trust are associated with the HDI, and whether the relationship between having high education and generalized trust varies according to a society's level of human development.

Education is measured in terms of an individual's highest degree, subsumed into three standard categories: low (highest degree below the upper-secondary level), intermediate (highest degree at the upper-secondary or nontertiary postsecondary level), and high (highest degree at the tertiary level). As with all categorical predictors, we include level of education using weighted effects (rather than dummy) coding. Weighted effects coding of categorical predictors is akin to grand mean

centering of continuous predictors. It ensures that the intercept corresponds to the predicted outcome for the average individual (te Grotenhuis et al. 2016). This not only eases interpretation, but it safeguards against problems that can arise in the estimation of mixed-effects models when the intercept corresponds to a highly idiosyncratic value near or even beyond the boundaries of the observed covariate distribution (Enders and Tofighi 2007). Given the weighted effects coding, the coefficient of the high-education indicator captures the (adjusted) difference in the respective outcome variable between high-educated individuals and individuals whose level of education equals that of the average European. Its cross-level interaction with the HDI indicates whether this difference changes with a society's level of human development.

In addition to the high-education indicator, the HDI, and (in half the specifications) their cross-level interaction, our models contain the following standard control variables: age (z-standardized around the grand mean) and (weighted effect-coded) indicators for having intermediate education, being female, being married, and being unemployed. Part A of the online supplement describes the predictor and outcome variables in more detail.

Varying Coefficients of Lower-Level Variables Are a Real Concern

Gender and at least some of the other five control variables are arguably included in most sociological regression models. If the associations between these six lower-level predictors and the five outcome variables vary considerably across ESS countries, we would thus have convincing evidence that the invariant coefficients assumption is dubious in many multilevel analyses. To investigate this possibility, we estimate, for each of the five outcome variables, a mixed-effects model that includes the HDI and all lower-level predictors. To examine the extent of cross-country heterogeneity in the effects of the lower-level variables, we allow the coefficients of all six predictors to vary across countries.

The results indicate substantial cross-cluster heterogeneity. To see this, consider the following best guesses of the country-specific coefficients based on the corresponding mixed-effects estimates (more technically, we report best linear unbiased predictions or BLUPs). Across Europe, the highly educated are less xenophobic than the average European by about .31 standard deviations. Yet, this average coefficient masks substantial cross-national differences. In Ukraine and Russia, being highly educated hardly makes a difference (b = –.09 and –.11, respectively). In Great Britain and Ireland, the effect is much more substantial (b = –.54 and –.50, respectively). Even the close link between education and occupational status varies across Europe. In Bulgaria and Lithuania, the highly educated hold jobs that are more than a standard deviation higher in status than those of individuals with an average level of education (b = 1.20 and 1.15, respectively). In Spain and the Netherlands, the advantage of holding a tertiary degree is smaller by about one third of a standard deviation (b = .88 and .91, respectively).

To give a comprehensive picture, we measure coefficient heterogeneity as the standard deviations of the random slopes, expressed in percent of the respective average coefficient (i.e., fixed effect). For each lower-level predictor, Figure 6 reports the median value of this measure across the five outcome variables (i.e., the ordered third value).[13] The two vertical lines indicate the two different values assumed in the Monte Carlo simulations.

Figure 6 shows that coefficient variation across clusters, as measured by the cross-cluster standard deviation, is typically at least half as large as the average coefficient. That is, within two standard deviations to the left and right of the average coefficient, there are countries where the control variable is hardly related to the outcome and others where the association is nearly twice as strong. In some cases (e.g., for the coefficient of being unemployed), median variability even exceeds 100 percent. Such high values might be the result of small average coefficients in the

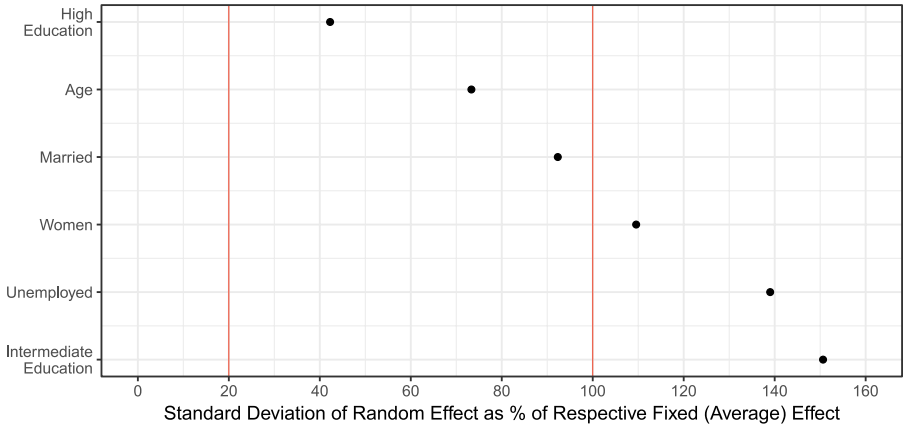

Figure 6. Cross-Country Variation in the Effects of Common Control Variables
Note: Cross-country variation is measured as the standard deviation of the random slope on a predictor, expressed in percent of the average effect (i.e., the fixed-effect coefficient estimate). The dots represent the median case (i.e., the third ordered value) across the five outcome variables studied in the illustrative analyses. The two vertical lines indicate the levels of variation assumed in the Monte Carlo simulations.

denominator. We therefore provide a more detailed figure that shows the average coefficient as well as the range of country-specific coefficients (again, based on BLUPs) for all 30 combinations of the five outcome and six predictor variables (see Figure E.6.1 in the online supplement). Although there are some small average coefficients, these clearly do not drive the results of Figure 6. In summary, this exploration of cross-country variation in the effects of standard controls not only validates our Monte Carlo simulations; it also carries the important message that such variation is likely a real concern in the typical (cross-country) multilevel analysis.

Maximally Flexible Models Do Not Perform Well in Practice

We have argued, and demonstrated through Monte Carlo simulations, that erroneously assuming the effects of lower-level variables to be invariant across clusters tends to reduce the precision of estimated context effects. One response to our results might be to simply estimate the maximally flexible model with random slopes for all lower-level variables and no constraints on their interrelations. However, as noted earlier, there are reasons to be concerned that such models might suffer from overparameterization, especially when the number of clusters is small.

We also discussed two-step estimation as an easy-to-implement alternative to mixed-effects models. It involves completely separate and unconstrained estimation of cluster-specific effects, which is advantageous when substantial heterogeneity exists, but also quite wasteful when there is no or very little actual heterogeneity. Our Monte Carlo results indicate that, given large lower-level samples, the benefits may outweigh the costs when the effects of at least some controls vary markedly.

To assess the performance of the different approaches under realistic conditions, we now turn to our illustrative empirical analyses. We measure the precision of estimated context effects using the width of two-sided 95 percent confidence intervals, focusing on three effects for each of the five outcome variables: the effect of the HDI from the DCE specification, and the cross-level interaction and main effect of the HDI from the CLI specification that includes an interaction term between the HDI and the high-education indicator. For each of these 15 effects of interest, we compare the confidence intervals from the maximally flexible models with

Table 3. Precision of Context Effect Estimates Relative to Invariant Specification: Maximally Flexible Mixed-Effects and Two-Step Model

		Maximally Flexible Mixed-Effects Model		Two-Step Model	
Outcome	Context Effect	ΔBootstrap CI	ΔAnalytic CI	ΔBootstrap CI	ΔAnalytic CI
		Direct Context Effect (DCE)			
Generalized trust	Direct HDI effect	–6.80	–33.71		
Homophobia	Direct HDI effect	.00	–23.46		
Xenophobia	Direct HDI effect	33.72	–16.01		
Fear of crime	Direct HDI effect	63.22	–17.63		
ISEI	Direct HDI effect	17.77	–19.88		
		Cross-Level Interaction (CLI)			
Generalized trust	Interaction effect	–4.57	–7.56	7.30	16.28
Homophobia	Interaction effect	17.41	–16.52	6.05	4.21
Xenophobia	Interaction effect	15.45	–18.48	.84	4.18
Fear of crime	Interaction effect	21.83	–6.72	13.82	4.11
ISEI	Interaction effect	–9.52	–13.56	32.20	25.71
Generalized trust	Main HDI effect	–22.71	–25.78		
Homophobia	Main HDI effect	–3.64	–25.33		
Xenophobia	Main HDI effect	12.63	–14.42		
Fear of crime	Main HDI effect	14.45	–12.79		
ISEI	Main HDI effect	–19.47	–22.84		

those from the invariant mixed-effects specification.[14]

Table 3 summarizes the results. We focus on differences Δ in the width of confidence intervals between the flexible models and the invariant one, expressed in percent of interval width for the invariant model.[15] Negative (positive) values thus mean that the given flexible model estimates the context effect more (less) precisely than the invariant model. Columns 3 (bootstrapped intervals) and 4 (analytic intervals) show how the maximally flexible mixed-effects model performs relative to the invariant one. Columns 5 (bootstrapped intervals) and 6 (analytic intervals) compare two-step estimation to the invariant mixed-effects model. We report comparisons based on analytic intervals for completeness, but we focus on the bootstrap-based comparisons because only the latter seem to provide accurate inference for complex mixed-effects models (see Figure 5). Table E.7.1 in the online supplement reports the interval widths underlying Table 3 and supports this result; bootstrapped confidence intervals for the maximally flexible mixed-effects model are systematically larger than their analytic counterparts, whereas no such pattern exists for the invariant model and the two-step estimates. This underlines that the bootstrapped intervals are preferable for comparisons of the different models.

Table 3 sends a rather sobering message about the performance of maximally flexible models. In only six of the 15 cases does the maximally flexible mixed-effects model outperform the invariant specification. In one case (DCE for homophobia), the difference is essentially zero. In the eight remaining cases, the invariant specification yields estimates that are substantially more precise. The most extreme case is the direct effect of the HDI on fear of crime, where the confidence interval for the maximally flexible specification is 63.2 percent wider than for the invariant one. Turning to two-step estimation, there is not a single case (out of five) where it produces more precise estimates than the invariant alternative.

In short, the evidence in Table 3 suggests that maximally flexible mixed-effects and two-step models will often be a suboptimal, and sometimes a very bad, choice. However, this does not mean practitioners should simply continue to use the invariant specification. The likely explanations for the sobering performance of the maximally flexible mixed-effects model also suggest a potential solution, namely to find an optimal balance between parsimony and flexibility by removing random effects that add little explanatory power.

Optimized Specifications Tend to Deliver Precision Gains

The task of balancing flexibility and parsimony is essentially one of model selection. As such, we can draw on the established toolbox of regression diagnostics and model selection techniques. That being said, simplification of the random-effects structure for mixed-effects models raises a few generic issues that many readers may not be familiar with. We therefore now describe the procedure we used to find more parsimonious random-effects specifications.

Our approach closely follows recent work by Bates, Kliegl, and colleagues (2015), hereafter BKVB, from a psycholinguistic context. The cases considered by BKVB tend to differ from the typical sociological application: for example, their data usually contain multiple, non-hierarchical levels of nesting. BKVB's overall goal is very similar to ours, however, namely to simplify the random-effects structure of a flexible mixed-effects model in a way that avoids overparameterization, while capturing the most important patterns of cross-cluster variation in the data. We provide a detailed example of the procedure in Part G of the online supplement. The replication files provide annotated step-by-step code for all analyses reported in Table 4 below.

The optimization routine starts with the maximally complex model and consists of two major and internally iterative steps. The first step seeks to iteratively reduce the number of random slopes by removing random effects that are not supported by the data. In each iteration, the crucial question is whether and how to simplify a given baseline specification. Initially, this baseline specification is the maximally flexible model with random slopes on all predictors. The simpler alternatives to the baseline specification, which we refer to as "candidates," comprise the specifications that result when one random slope is dropped at a time.[16] Following BKVB, it is useful to speed up estimation by constraining all correlations to zero during this step. Two criteria determine if one of the (simpler) candidates is preferable to the (more complex) baseline specification: (1) changes in BIC (other selection criteria are possible, but we found BIC to perform best[17]) and (2) the results of a principal component analysis (PCA). As for (1), lower BIC values are preferable, so the most promising alternative to the baseline specification, the best candidate, is the one with the lowest BIC value. If the BIC value of the best candidate is lower than that of the baseline specification, the former is preferable and becomes the new baseline specification for the next iteration. If the BIC value of the best candidate exceeds that of the baseline specification, we turn to (2) and the decision depends on the results of the PCA. The goal of the PCA is to find out whether the data provide sufficient information to support the complexity of the (baseline) model.[18] BKVB argue persuasively that the number of principal components that cumulatively account for 100 percent of the variance of the random effects is an upper limit for the number of random effects supported by the data. Thus, if the PCA indicates that the number of principal components that jointly account for the random-effects structure is smaller than the number of random effects, we prefer the best candidate to the baseline specification even if it has a higher BIC than the latter. The best candidate then becomes the new baseline specification in the next iteration of the first step. If the PCA indicates no need for further simplification and the best candidate has a higher BIC than the baseline specification, the latter is preferable and the first step of the optimization routine is complete.

Once the first step of the procedure has been completed, the second step seeks to simplify the structure of the correlations among them. Again, this step is iterative and consists of repeated comparisons between a baseline specification and one or several candidates with a simpler correlation structure. Initially, the baseline specification is the model with all random effects remaining after the first step and no constraints on the correlations among them. After the first iteration, it is the best candidate from the previous step. Identification of candidates is somewhat more difficult than in the first step of the routine, because the number of possible correlation structures is quite large unless the number of random slopes is small. It is therefore useful to examine the random-effects covariance matrix of the baseline specification for low correlations and to identify a small subset of promising candidates. Again, the choice between the baseline specification and the candidates should then be based on a transparent criterion (in our case, BIC). The second step of the procedure terminates when no further improvements are possible.

One complication needs to be taken into account during this second step of the optimization routine. Constraining the correlation between the random slope for a lower-level predictor, say x, and the random intercept to zero can render coefficient estimates on contextual variables sensitive to the scaling of x (Bates, Mächler, Bolker, and Walker 2015). Centering of continuous and (weighted) effect coding of categorical predictors (te Grotenhuis et al. 2016) help safeguard against this issue. Nevertheless, researchers should verify that estimated context-effect sizes are robust to the removal of correlation terms, while noting that some variation is to be expected simply due to chance and changes in precision. If a context effect changes substantially, one should either (re)introduce the respective correlation (i.e., use a more complex model) or remove the uncorrelated random slope entirely.

Do optimized models produce more precise estimates of context effects than the invariant specification? Table 4 shows they typically do. As in Table 3, we report relative differences in interval width between the optimized and invariant specifications. Focusing on the bootstrap-based comparisons in the penultimate column, we now find that for six of the 15 context effects, using a more flexible (optimized) specification increases precision by over 10 percent. For another four effects, we find modest gains between 2 and 4 percent. For four of the remaining five effects, the difference is between −2 and +1 percent. In only one case, the cross-level interaction between HDI and the high-education indicator in the model for fear of crime, do we find the optimized model to produce a noticeably less precise estimate than the invariant specification, with the difference being approximately 4.4 percent. Table 4 (see columns three and four) shows that flexible (optimized) models need to be much more parsimonious than the maximally flexible model to achieve these gains in accuracy. Whereas the maximally flexible model specifies seven random effects (one intercept and six slopes) as well as 21 correlations among these, the optimized models retain between two and four random slopes. Moreover, several correlations are usually constrained to zero (the maximum number of correlations varies across models because it depends on the number of random slopes that remain after the first optimization step).

Table E.7.1 in the online supplement reports the absolute widths of the confidence intervals that underlie the relative comparisons in Table 4. Interestingly, it shows that bootstrapped confidence intervals for the optimized flexible models are not systematically larger than their analytic counterparts. As with the invariant model, the bootstrapped confidence intervals are generally of broadly similar magnitude. This is consistent with overparameterization being the primary reason why analytic confidence intervals for ME-Correct showed severe undercoverage in the simulations, particularly in experimental conditions with few clusters or many varying effects (cf. Figures 3 and 5). It suggests that model optimization can greatly reduce or perhaps even resolve the problems with conventional analytic inference. To further examine

Table 4. Precision of Context Effect Estimates Relative to Invariant Specification: Optimized Mixed-Effects Model

Outcome	Context Effect	Remaining Random Effects	Remaining Correlations	ΔBootstrap CI	ΔAnalytic CI
		Direct Context Effect (DCE)			
Generalized trust	Direct HDI effect	4/7	1/6	−2.24	−12.54
Homophobia	Direct HDI effect	5/7	2/10	.24	−.63
Xenophobia	Direct HDI effect	5/7	2/10	−13.85	−11.70
Fear of crime	Direct HDI effect	4/7	1/6	−15.08	−10.29
ISEI	Direct HDI effect	5/7	2/10	−11.40	−9.53
		Cross-Level Interaction (CLI)			
Generalized trust	Interaction effect	3/7	1/3	−3.83	1.79
Homophobia	Interaction effect	5/7	4/10	−.04	−13.70
Xenophobia	Interaction effect	5/7	4/10	−2.46	−15.49
Fear of crime	Interaction effect	4/7	1/6	4.36	−3.97
ISEI	Interaction effect	5/7	6/10	−2.19	−7.71
Generalized trust	Main HDI effect	3/7	1/3	.48	.17
Homophobia	Main HDI effect	5/7	4/10	−1.45	−6.42
Xenophobia	Main HDI effect	5/7	4/10	−10.40	−10.75
Fear of crime	Main HDI effect	4/7	1/6	−11.51	−8.96
ISEI	Main HDI effect	5/7	6/10	−26.36	−21.95

this possibility, we ran another set of Monte Carlo trials, evaluating the coverage rate of analytic confidence intervals for specifications optimized according to the above procedure. We did so for the same experimental condition used to evaluate the performance of the non-parametric bootstrap (15 countries, moderate compositional differences [15 percent of variance between clusters], three controls with varying slopes, each with a standard deviation of 1).[19] The actual coverage rate of analytic confidence intervals for the optimized specifications across 10,000 Monte Carlo trials was 92.00 (cf. the rightmost bar in Figure 5). This is a substantial improvement over the coverage rate of analytic confidence intervals for ME-Correct (85.67 percent), but it still falls short of the nominal level. Until this issue has been examined more comprehensively, we recommend that researchers base their final inference on bootstrapped confidence intervals.

Overall, our illustrative analyses of ESS data suggest that eschewing the predominant invariant model for a more flexible specification can substantially improve the precision of estimated context effects in real-life settings. But to make the most of their data, researchers need to find the right balance between complexity and parsimony. There may of course be room for improvement, but our slightly adapted version of BKVB's optimization procedure provides a transparent and principled algorithm for achieving this goal.

CONCLUSIONS

Thanks to a long-standing interest in context effects, ever-improving data availability, and methodological advances, quantitative analysis of multilevel data has become a staple of applied work in sociology. Our review of current multilevel modeling practices in leading sociology journals shows that analysts rarely allow the coefficients of lower-level controls to vary across clusters, even when they deal with very heterogeneous ones such as countries. Based on semi-formal reasoning, we argued that this "invariant coefficients assumption" leads to biased estimates of within-cluster relationships and thereby jeopardizes the reliability of estimated context effects. Illustrative multilevel analyses using data on 28 countries from the European

Social Survey demonstrate that varying coefficients are a real concern: for five outcome variables, we found that cross-country variation in the effects of standard controls, such as gender, age, and level of education, is typically substantial.

We explored the consequences of the invariant coefficients assumption using Monte Carlo simulations. Our results show that erroneously assuming the coefficients of lower-level predictors to be invariant across clusters can markedly reduce the precision of parameter estimates for both direct context effects and cross-level interactions. By the same token, models that do allow for varying effects achieve greater precision and statistical power, that is, they have better chances of rejecting the null hypotheses of no effect when it is in fact wrong. The Monte Carlo analysis shows that the consequences of neglecting cross-cluster heterogeneity are particularly severe when the variation of coefficients is substantial (i.e., when there are many lower-level variables whose effects vary or when the variability of effects is high), when clusters differ markedly in terms of their composition with respect to lower-level variables (direct context effects only), and when there are few clusters. These conditions are typical of country-comparative studies, where concerns about precision are also most salient because so few cases are available for identifying the contextual effects of interest.

OLS with conventional cluster-robust standard errors delivered the poorest performance in the Monte Carlo simulations, in terms of both precision and inference. Two-step estimation generally produced accurate statistical inference. It also achieved relatively high precision when clusters were large (as they typically are in country comparisons) and when there was substantial heterogeneity in the effects of controls. The predominant model in applied sociological research, a mixed model assuming the effects of all lower-level control variables to be invariant across clusters, achieved good results overall, generally producing slightly more precise estimates of context effects than did OLS with cluster-robust standard errors. In contrast to some recent analyses (Bryan and Jenkins 2016; Stegmueller 2013), our analysis shows that it provides accurate statistical inference for context effects when certain guidelines are followed (for details, see Elff et al. 2016).

The most important result of the Monte Carlo analysis, however, is that a more flexible mixed-effects specification with random slopes on lower-level variables whose effects actually vary across clusters produced more precise estimates of context effects than did the invariant alternative. A drawback of flexible mixed-effects models is that conventional statistical inference tends to be anticonservative: analytic confidence intervals are too narrow and *p*-values too optimistic. Fortunately, additional simulation evidence suggests that this problem can be overcome by using a non-parametric cluster bootstrap. Moreover, our empirical illustrations and further Monte Carlo analysis suggest that overparameterization is the primary reason why conventional analytic inference for complex models falls short. Careful model selection should therefore attenuate the problem, but to be on the safe side we recommend that practitioners use bootstrapping for their final inferences.

Our empirical illustrations using ESS data are consistent with the simulation results and suggest that practitioners would be well-advised to "[c]onsider all coefficients as potentially varying" (Gelman and Hill 2007:549). Yet, this does not mean one should blindly assume that all coefficients vary. Small upper-level samples will often provide insufficient information for estimating the maximally flexible mixed-effects specification with random slopes on all lower-level controls. Analysts therefore need to find a random-effects specification that captures the most important aspects of cross-cluster heterogeneity while avoiding overparameterization. This task of model selection is not unique to the present setting, and many practitioners will already be proficient in it. In particular, one can rely on several well-established tools, such as regression diagnostics and information criteria in the search for an optimal specification. We have outlined, and examined the performance of, an

optimization procedure based on recent work by Bates, Kliegl, and colleagues (2015). The method uses standard tools for model comparison—in our case, BIC worked best—but also involves a check that is more generic to the task at hand, namely that of subjecting the random-effects structure to a principal component analysis to identify overparameterization. The resulting optimized specifications compare rather favorably with the invariant alternative that predominates in applied research. In our illustrative analyses, the increase in precision achieved by the optimized models is more modest than in the Monte Carlo simulations. Nevertheless, achieving gains of comparable magnitude (e.g., a 10 percent increase) without improving the model would require substantially larger samples at the cluster level—something that practitioners frequently yearn for.

Putting all our arguments and findings in a nutshell, we make the following recommendations: consider all coefficients as potentially varying, but find the right balance between flexibility and parsimony by using a principled optimization routine, and finally bootstrap the standard errors to be on the safe side.

Despite its promising results and apparently clear message, our study has limitations that should be addressed in future research. The outlined strategy for model selection may not yet be the ideal one. The resulting models often achieve substantial precision gains, but they do not do so in all illustrative analyses. Furthermore, even though imposing constraints on the correlations among random effects is an effective means of simplifying the random-effects structure, it introduces a risk that estimated context effects become sensitive to the scaling of lower-level controls. Centering and weighted effects coding of predictors alleviate this problem, yet careful analysis remains imperative. It is also important to note that the primary criterion in the strategy outlined here is overall model fit. Yet, an ideal selection strategy for the kinds of research questions we focused on should perhaps put special emphasis on the estimation of context effects. Our analysis suggests several aspects that might be central in such an approach: for example, varying coefficients on a lower-level predictor should not necessarily matter much if that predictor shows little cross-cluster compositional differences (direct-context-effects case) or if it is hardly correlated with the lower-level variables whose varying coefficients we seek to explain using cross-level interactions. Whatever the approach taken, we believe that researchers should provide a clear rationale for their specification choices. Indeed, one benefit of a transparent and principled algorithm is that it reduces the risk of cherry-picking (e.g., arbitrarily choosing specifications that produce particularly strong effects).

Further obvious questions for future research are how the different approaches perform when the data have a more complex structure and when heterogeneity in the effects of controls takes more pernicious forms. For example, many cross-national surveys such as the ESS now cover several waves, leading to the diffusion of more complex multilevel models with multiple (cross-classified) levels of nesting (Schmidt-Catran and Fairbrother 2016). It may be worthwhile to consider situations where cluster differences in the effects of controls are systematically related to a context variable of interest. Another possible extension is to repeat the analysis for non-continuous outcomes, although we would expect results to be qualitatively similar.

The central insight from our study is that heterogeneity in the effects of lower-level variables is a crucial possibility to consider in multilevel settings. Our semi-formal reasoning, Monte Carlo simulation evidence, and several illustrative multilevel analyses all demonstrate that, even in its most benign forms, it can depress the ability of predominant estimators to detect and precisely quantify context effects. Questions about context effects rank among the most important ones that sociologists study, yet they are often difficult to answer because so few clusters are available—as, for example, in country comparisons that seek to identify policy effects. Our analysis suggests that we need to reassess our empirical strategies to ensure we make optimal use of the potential offered by multilevel data.

Acknowledgments

Parts of this paper were presented at the 2013 and 2015 Conferences of the European Survey Research Association, the XVIII International Sociological Association World Congress of Sociology, the RC 28 Spring Meeting 2016, and in seminars at Humboldt University Berlin, WZB Berlin Social Science Center, Goethe University Frankfurt, the University of Amsterdam, and Ludwig-Maximilians University Munich. We thank participants for their useful feedback. We are particularly indebted to Felix Elwert, Lena Hipp, Ulrich Kohler, Joscha Legewie, and the *ASR* editors and anonymous reviewers for numerous helpful comments and suggestions.

Notes

1. We did not count analyses of vignettes (factorial surveys) or panel analyses as multilevel studies, although such data also have a clustered structure and are often analyzed using mixed-effects models.
2. Even counting sets of dummies as one variable, the average study still included 8.41 lower-level predictors.
3. For the derivation, see Part B of the online supplement.
4. The lower-level predictors x_{1ig} to x_{6ig} are multivariate normal with standard deviations of one, and correlation matrix Σ. In the setting with no compositional differences (see below) all predictors have means of zero. To ensure that our findings do not hinge on the (arbitrary) choice of a specific correlation matrix, each simulation run generates Σ randomly using a generalization of the algorithm proposed by Joe (2006). The pairwise correlations in Σ are positive and negative with equal probability. Their average strength (in absolute terms) is .3.
5. Because the random effects are (multivariate) normal, the middle 90 percent of cluster-specific effects fall into the range .67 to 1.33 when we set the standard deviation to .2. When we set it to 1, the 90 percent range runs from –.64 to 2.64. As with the lower-level predictors, we generate random correlation matrices using a generalization of the algorithm by Joe (2006). The average absolute correlation between the random effects is .33. In Part E of the online supplement we show simulation results based on a lower average absolute correlation of .2.
6. Technically, we implement compositional differences by adding random cluster-specific constants to the lower-level predictors. The constants are uncorrelated across predictors and normally distributed with means of zero. We control the extent of compositional differences by changing the standard deviation of the cluster-specific constants.
7. This approach seeks to improve precision by weighting down unreliable estimates in the cluster-level regression. Additional analyses show that using OLS in both steps yields similar results.
8. Further simulations indeed show that the RMSE of ME-Correct hardly declines at all with the number of random slopes when we use a DGP with lower correlations among the random effects (see Figures E.1.1 and E.1.2 in the online supplement).
9. The one qualification to this statement is that in the country scenarios high levels of heterogeneity (cross-cluster standard deviation of 1) appear to exacerbate coverage when all five lower-level controls have varying effects.
10. We calculated this by multiplying the RMSE of ME-Correct (16.4 percent) with the 90th percentile of the normal distribution (ca. 1.282).
11. As before, actual coverage rates of analytic confidence intervals are based on 10,000 replications.
12. Except for Kosovo where we detected problems with one of the lower-level variables (marital status), we include all countries that participated in round 6 of the ESS.
13. Additional analyses (available upon request) show that the extent of cross-national heterogeneity in the coefficients of the six predictors is broadly similar if we calculate it using mixed-effects models that constrain the correlations between the random effects to zero or using country-specific linear regressions (i.e., the first step of two-step estimation).
14. The invariant specification includes only a random intercept in the DCE case and a random intercept as well as a random slope on being highly educated in the CLI case. The maximally flexible mixed-effects model includes a random intercept and random slopes on all six lower-level predictors in both cases. The covariances between the random effects are not constrained and estimated from the data. The two-step estimates are based on running country-specific regressions in a first step and then regressing the coefficient estimates for the high-education indicator on the HDI (as before, we consider this approach only for the cross-level interaction term).
15. The measure is directly comparable to the reduction in RMSE (or standard error) emphasized in the discussion of the Monte Carlo evidence. To see this, note that a standard two-sided 95 percent confidence interval based on the normal approximation has a width of approximately ($2 \times 1.96 \times$ SE). It follows immediately that reducing the SE by a constant factor reduces the width of the confidence interval by the same factor.
16. We never remove the random slope on a lower-level predictor that is part of a cross-level interaction, because the inclusion of this random slope is crucial for correct inference.
17. BKVB focus on likelihood ratio tests, taking an insignificant test result as evidence that the simpler specification is preferable. In our case, this criterion is of little use because, given our sample size of roughly 50,000 individuals, likelihood ratio tests practically always favor the more complex specification. Another obvious option would be to consider changes in AIC. AIC tends to penalize additional parameters less harshly than BIC, by a factor of 2 rather than $\log(n)$, where n is the total sample size

(Müller, Scealy, and Welsh 2013). In fact, we also conducted the optimization routine using the AIC criterion. Overall, the resulting models performed worse than those based on BIC (results available upon request), so we report only the latter ones in the interest of brevity.

18. In R, the PCA of the random-effects matrix is implemented in the *rePCA* function of the *RePsychLing* package available at (https://github.com/dmbates/RePsychLing). We have written a Stata program that implements the method for two-level hierarchical models. It is available for download with the online supplements.
19. To conduct these simulations, we had to automate the optimization procedure. To simplify and speed up the process, we had the algorithm start from the random-effects specification for ME-Correct, that is, from a specification with a random intercept and random slopes on the three lower-level predictors whose effects actually vary in the DGP for the experimental condition. The effects of the remaining lower-level variables were treated as fixed from the outset. We thus used information that would not be available in an actual application, but the random slopes on the non-varying predictors would usually have been pruned during the optimization procedure anyway. The algorithm then proceeded with the first step of the optimization procedure as described earlier, eliminating or retaining random slope terms based on changes in BIC and the results of the PCA. In the second optimization step (elimination of correlations), the algorithm considered all possible specifications and chose the one with the lowest BIC value.

References

Bates, Douglas, Reinhold Kliegl, Shravan Vasishth, and Harald Baayen. 2015. "Parsimonious Mixed Models." *arXiv* preprint arXiv:1506.04967. Retrieved April 22, 2016 (http://arxiv.org/abs/1506.04967).

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67(1):1–48.

Bryan, Mark L., and Stephen P. Jenkins. 2016. "Multilevel Modelling of Country Effects: A Cautionary Tale." *European Sociological Review* 32(1):3–22.

Cameron, A. Colin, and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50(2):317–72.

Durkheim, Emile. 1897. *Suicide: A Study in Sociology*. Glencoe, IL: Free Press.

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." *SocArXiv/Open Science Framework* (Version 2, December 10, 2016; https://osf.io/preprints/socarxiv/z65s4/).

Enders, Craig K., and Davood Tofighi. 2007. "Centering Predictor Variables in Cross-Sectional Multilevel Models: A New Look at an Old Issue." *Psychological Methods* 12(2):121–38.

Esarey, Justin, and Andrew Menger. 2015. "Practical and Effective Approaches to Dealing with Clustered Data." Presented at the 2015 Annual Meeting of the Society for Political Methodology at the University of Rochester, Rochester, New York.

European Social Survey (ESS) Round 6. 2016. *ESS-6 2012 Documentation Report*, ed. 2.2. Bergen: European Social Survey Data Archive, NSD – Norwegian Centre for Research Data for ESS ERIC.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchichal Models*. Cambridge, UK: Cambridge University Press.

Goldstein, Harvey. 2011. "Bootstrapping in Multilevel Models." Pp. 163–71 in *Handbook of Advanced Multilevel Analysis*, edited by J. J. Hox and J. K. Roberts. New York: Routledge.

Imbens, Guido W., and Michal Kolesár. 2016. "Robust Standard Errors in Small Samples: Some Practical Advice." *Review of Economics and Statistics* 98(4):701–712.

Joe, Harry. 2006. "Generating Random Correlation Matrices Based on Partial Correlations." *Journal of Multivariate Analysis* 97(10):2177–89.

Kézdi, Gabor. 2004. "Robust Standard Error Estimation in Fixed-Effects Panel Models." *Hungarian Statistical Review* 89(9):96–116.

Kloek, Teun. 1981. "OLS Estimation in a Model Where a Microvariable Is Explained by Aggregates and Contemporaneous Disturbances Are Equicorrelated." *Econometrica* 49(1):205–207.

Lewis, Jeffrey B., and Drew A. Linzer. 2005. "Estimating Regression Models in Which the Dependent Variable Is Based on Estimates." *Political Analysis* 13(4):345–64.

Lindley, Dennis V., and Adrian F. M. Smith. 1972. "Bayes Estimates for the Linear Model." *Journal of the Royal Statistical Society. Series B (Methodological)* 34(1):1–41.

Maas, Cora J. M., and Joop J. Hox. 2004. "The Influence of Violations of Assumptions on Multilevel Parameter Estimates and Their Standard Errors." *Computational Statistics & Data Analysis* 46(3):427–40.

MacKinnon, James G., and Matthew D. Webb. 2014. "Wild Bootstrap Inference for Wildly Different Cluster Sizes." Working paper, Queen's Economics Department, Queen's University, Kingston, Canada. Retrieved June 26, 2015 (http://www.econstor.eu/handle/10419/97471).

Moulton, Brent R. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics* 32(3):385–97.

Moulton, Brent R. 1990. "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units." *Review of Economics and Statistics* 72(2):334–38.

Müller, Samuel, J. L. Scealy, and A. H. Welsh. 2013. "Model Selection in Linear Mixed Models." *Statistical Science* 28(2):135–67.

Raudenbush, Stephen W., and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd ed. Thousand Oaks, CA: Sage Publications.

Rogers, William. 1993. "Regression Standard Errors in Clustered Samples." *Stata Technical Bulletin* 3(13):19–23.

Sampson, Robert J. 2013. *Great American City: Chicago and the Enduring Neighborhood Effect*. Chicago: University of Chicago Press.

Schmidt-Catran, Alexander W., and Malcolm Fairbrother. 2016. "The Random Effects in Multilevel Models: Getting Them Wrong and Getting Them Right." *European Sociological Review* 32(1):23–38.

Stegmueller, Daniel. 2013. "How Many Countries for Multilevel Modeling? A Comparison of Frequentist and Bayesian Approaches." *American Journal of Political Science* 57(3):748–61.

te Grotenhuis, Manfred, Ben Pelzer, Rob Eisinga, Rense Nieuwenhuis, Alexander Schmidt-Catran, and Ruben Konig. 2016. "When Size Matters: Advantages of Weighted Effect Coding in Observational Studies." *International Journal of Public Health* 62(1):163–67.

United Nations Development Programme, ed. 2015. *Human Development Report 2015*. New York: United Nations Development Programme.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48(4):817–38.

Williams, Rick L. 2000. "A Note on Robust Variance Estimation for Cluster-Correlated Data." *Biometrics* 56(2):645–46.

Wooldridge, Jeffrey M. 2003. "Cluster-Sample Methods in Applied Econometrics." *American Economic Review* 93(2):133–38.

Wooldridge, Jeffrey M. 2014. *Introduction to Econometrics. Europe, Middle East and Africa Edition*. Andover, MA: Cengage Learning.

Jan Paul Heisig is a senior researcher in the research unit Skill Formation and Labor Markets at WZB Berlin Social Science Center. His interests include social inequality, labor market dynamics, education, and quantitative methods. Recent articles have appeared in *Ageing & Society, Research in Social Stratification and Mobility*, and *Sociology of Education*. His monograph *Late-career Risks in Changing Welfare States* was published with Amsterdam University Press in 2015.

Merlin Schaeffer is Professor of Demography and Social Inequality at University of Cologne. His research interests include the comparative analysis of population dynamics and social stratification, as well as quantitative methodology. His recent research projects focus on the labor market consequences of ability-qualification mismatches among persons of immigrant origin, and the role of contextual-demographic characteristics for inter-ethnic relations. His recent work has appeared in *American Journal of Sociology, European Sociological Review*, and *Social Science Research*.

Johannes Giesecke is Professor of Empirical Social Research at Humboldt University Berlin. His research topics are social inequality, labor market sociology, migration and integration, and quantitative methods. His work has been published in *European Sociological Review, Research in Social Stratification and Mobility*, and *Social Forces*.