European Sociological Review, 2019, 1–22
doi: 10.1093/esr/jcy053
Original Article

Why You Should *Always* Include a Random Slope for the Lower-Level Variable Involved in a Cross-Level Interaction

Jan Paul Heisig [1,†] **and Merlin Schaeffer** [2,*,†]

[1]WZB Berlin Social Science Center, Reichpietschufer 50, 10785 Berlin, Germany and [2]University of Copenhagen, Øster Farimagsgade 5, DK-1353 København, Denmark

*Corresponding author. Email: mesc@soc.ku.dk

†Both authors have contributed equally.

Submitted January 2018; revised December 2018; accepted December 2018

Abstract

Mixed-effects multilevel models are often used to investigate cross-level interactions, a specific type of context effect that may be understood as an upper-level variable moderating the association between a lower-level predictor and the outcome. We argue that multilevel models involving cross-level interactions should always include random slopes on the lower-level components of those interactions. Failure to do so will usually result in severely anti-conservative statistical inference. We illustrate the problem with extensive Monte Carlo simulations and examine its practical relevance by studying 30 prototypical cross-level interactions with European Social Survey data for 28 countries. In these empirical applications, introducing a random slope term reduces the absolute *t*-ratio of the cross-level interaction term by 31 per cent or more in three quarters of cases, with an average reduction of 42 per cent. Many practitioners seem to be unaware of these issues. Roughly half of the cross-level interaction estimates published in the *European Sociological Review* between 2011 and 2016 are based on models that omit the crucial random slope term. Detailed analysis of the associated test statistics suggests that many of the estimates would not reach conventional thresholds for statistical significance in correctly specified models that include the random slope. This raises the question how much robust evidence of cross-level interactions sociology has actually produced over the past decades.

Introduction

One of the enduring questions of sociology is how human attitudes and behaviour are shaped by the social environment and how *vice versa* the social environment emerges from human action. The investigation of context effects, where an environmental feature (e.g., a characteristic of a neighbourhood or country) affects processes at a lower level (e.g., that of the individual), is therefore central to the discipline, and one should think that sociologists are highly proficient in modelling them statistically.

Quantitative sociologists typically use mixed-effects models, which are also known as ‘hierarchical models’ or simply ‘multilevel models’, to deal with the statistical challenges that arise in the estimation of context effects (see the ‘Mixed Effects Models with Cross-Level Interactions’ section and Equations 1–4 below). A crucial issue in the specification of these models is the choice of a random-effects structure (i.e., random intercept and slopes), which can have important consequences both for the precision of parameter estimates (Heisig, Schaeffer and Giesecke, 2017) and for statistical inference (Berkhof and Kampen, 2004; Barr *et al.*, 2013; Bryan and Jenkins, 2016; Schmidt-Catran and Fairbrother, 2016; Bell, Fairbrother and Jones, 2018).

The random-effects structure is also a crucial issue in the estimation of cross-level interactions, which are a special type of context effect where a contextual characteristic moderates the strength of a lower-level relationship (see Equation 4 below). To fix ideas, consider the following example, which also serves as one of the illustrative empirical examples presented later on: The (individual-level) relationship between fear of crime (as the outcome) and education (as the predictor) might be weaker in less developed countries (as indicated by the human development index; HDI) where the generally poor living conditions instil a fear of crime into everyone. Or to put it another way, the better-educated tend to benefit the most from improving societal conditions, whereas the less educated continue to live in fear of crime even in more developed societies.

Researchers who study cross-level interactions are interested in variation of lower-level relationships across contexts. One might therefore expect their models to include so-called random slope terms that capture unexplained contextual variation in these relationships (see Equation 3 below for a formal representation). In our example, one would include a random slope to account for cross-country differences in the relationship between education and fear of crime that are not explained by country differences in human development.

A review of published research, however, reveals that in many analyses of cross-level interactions the corresponding random slope is missing. Between 2011 and 2016, the *European Sociological Review* (*ESR*) published 28 studies that investigated cross-level interactions using (two-level) mixed-effects multilevel models (24 of these studies were country comparisons). More than half of these studies (17/28 or 61 per cent) only specified random intercept models without any random slopes (for details, see the ‘Cross-Level Interactions in the *ESR*’ section).

Given that empirical practice is so inconsistent, one may wonder whether the inclusion of random slope terms on the lower-level components of cross-level interactions is a matter of taste or whether one approach will usually be preferable to the other. A review of prominent textbooks on multilevel modelling does not provide a clear answer. In one widely read book, Snijders and Bosker (2012) note that ‘tested fixed effects’ should be accompanied by ‘an appropriate error term [. . .] For cross-level interactions, it is the random slope of the level-one [i.e., lower-level] variable involved in the interaction’ (p. 104). Other authors take a more ambiguous position. For example, Raudenbush and Bryk’s (2002) book includes a section on ‘A Model with Nonrandomly Varying Slopes’ where they suggest that a model with a cross-level interaction may omit the corresponding random slope if ‘little or no variance in the slopes remains to be explained’ (p. 28). They provide no precise definition of ‘little or no variance’, however. In their chapter on ‘Random-coefficient models’, Rabe-Hesketh and Skrondal (2012) generally include random slope terms alongside cross-level interactions, but they also note that the decision whether to do so often seems to be driven by technicalities of the software used: ‘Papers using HLM tend to include more cross-level interactions and more random coefficients in the models (because the level-2 [i.e., upper-level] models look odd without residuals) than papers using, for instance, Stata’ (p. 212f.). This certainly does not sound like an emphatic recommendation to include the random slope for statistical reasons.

In this article, we argue that such a recommendation should be given. We explain and demonstrate that the omission of random slopes in the analysis of cross-level interactions constitutes a specification error that will often have severe consequences for statistical inference about the coefficient of the cross-level interaction term (i.e., in our running example, the interaction between education and HDI) and about the main effect of the lower-level predictor involved in the interaction (i.e., the main effect of education). Only the main effect of the upper-level predictor remains unaffected (provided that the model includes a random intercept, as is generally the case in applied research).

In the next section, we briefly introduce mixed-effects models with cross-level interactions. In the ‘Why *Always* a Random Slope?’ section, we then explain that random slopes capture cluster-driven heteroskedasticity and cluster-correlated errors. As in standard linear regression, ignoring heteroskedasticity and within-cluster error correlation by failing to specify the appropriate random slope term will typically lead to downward bias in standard error estimates.

The two subsequent sections present Monte Carlo simulations and illustrative empirical analyses that support our claims. The simulations show that (correctly specified) mixed-effects models with a random intercept and a random slope on the lower-level component of the cross-level interaction generally achieve accurate statistical inference for all coefficients of interest. By contrast, random intercept models that omit the random slope term produce severely anti-conservative inference for the cross-level interaction term and the main effect of its lower-level component. The proportion of 95 per cent confidence intervals that do not cover the true effect size (i.e., the actual coverage rate) is generally smaller than the nominal rate, and often by a substantial margin. We find that the extent of undercoverage increases with the extent of variation in the (unmodelled) random slope, the variance of the lower-level component, and the number of lower-level observations per cluster. Illustrative empirical analyses of European Social Survey (ESS) data for 28 countries indicate that the consequences of omitting the random slope on the lower-level component are severe in real-life settings. We examine a total of 30 cross-level interactions and find that inclusion of the random slope term deflates the absolute *t*-ratio on the cross-level interaction term by 31 per cent or more in three quarters of cases, with an average reduction of 42 per cent.

We then review studies of cross-level interactions published in the *ESR* between 2011 and 2016. Unsurprisingly, we find that authors were more likely to report statistically significant cross-level interactions when they used a misspecified model that omitted the corresponding random slope. Consistent with '*P*-hacking' (Simonsohn, Nelson and Simmons, 2014), the distribution of absolute *t*-ratios for models estimated without a random slope exhibits a marked peak just above the critical value of 1.96. In combination with the results of our Monte Carlo simulations and empirical illustrations, our review therefore suggests that many published estimates based on models omitting the random slope would not have reached conventional levels of statistical significance in a correctly specified model.

The subsequent and penultimate section presents a further result of our analysis: the omission of a relevant random slope also leads to anti-conservative inference for a corresponding 'pure' lower-level effect. That is, even if the model does not contain any cross-level interactions involving education, accurate inference for the average effect of education on fear of crime across the 28 ESS countries would require the inclusion of a random slope on education—provided that such a slope is present in the process that gave rise to the data. While this result is troubling, there are two reasons to be less concerned than in the cross-level interaction case. First, most sociologists who use multilevel models are primarily interested in context effects rather than pure lower-level effects, as we confirm through a systematic analysis of the titles, abstracts, and formal hypotheses of research published in the *ESR*. Second, pure lower-level effects can typically be estimated with much greater precision (and correspondingly higher absolute *t*-statistics) than cross-level interactions. As a consequence, estimated lower-level effects should often stay statistically highly significant even if the associated *t*-ratio declines by 50 per cent or more. In the cross-level interaction case, such a decrease will often mean the difference between moderately strong and no statistically meaningful evidence against the null hypothesis.

The concluding section discusses the primary implications of our study. Looking backward, our findings suggest that the empirical basis for many seemingly well-established findings in comparative research may be much shakier than previously thought. Looking forward, a minimum requirement for future studies that examine cross-level interactions using multilevel models is to include a random slope on the corresponding lower-level variable. However, our findings suggest that fully accurate statistical inference for *all* coefficients, including pure lower-level effects, requires the inclusion of additional random slopes or alternative methods of inference, an important issue that should be addressed in future work.

Mixed-Effects Models with Cross-Level Interactions

In a first step, we briefly review the general logic of mixed-effects models with cross-level interactions (for comprehensive introductions, see, for example, Raudenbush and Bryk, 2002; Rabe-Hesketh and Skrondal, 2012; Snijders and Bosker, 2012). We begin with the following lower-level equation for the (lower-level) outcome Y_{ij} (e.g., fear of crime):

$$Y_{ij} = \beta_j^{(c)} + \beta_j^{(x)} x_{ij} + \epsilon_{ij}, \qquad (1)$$

where i indexes lower-level observations (e.g., individuals) and j indexes upper-level observations or clusters (e.g., countries). $\beta_j^{(c)}$ is the constant (i.e., intercept) and $\beta_j^{(x)}$ is the coefficient of lower-level predictor x_{ij} (e.g., education). The subscript j on the two parameters, $\beta_j^{(c)}$ and $\beta_j^{(x)}$, indicates that both are considered as potentially varying across clusters. In terms of our example, the j on $\beta_j^{(x)}$ thus means that the degree to which better-educated people are less afraid of crime might vary across countries. The model could be extended to include additional lower-level predictors x_{2ij} to x_{kij}, but for our analysis, this is not necessary. ϵ_{ij} is a lower-level error often assumed to follow $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, that is, to be normally distributed with a mean of zero and constant variance σ^2 (homoskedasticity).

In a cross-level interaction model, $\beta_j^{(x)}$ is specified as dependent on at least one cluster-level (i.e., contextual) variable z_j (e.g., the HDI). Typically, the model will (and should) also allow for a relationship between the constant $\beta_j^{(c)}$ and z_j. One way to formalize this is to write $\beta_j^{(c)}$ and $\beta_j^{(x)}$ as the outcome variables in two cluster-level equations:

$$\beta_j^{(c)} = \gamma^{(c)} + \gamma^{(cz)} z_j + u_j^{(c)} \tag{2}$$

and

$$\beta_j^{(x)} = \gamma^{(x)} + \gamma^{(xz)} z_j + u_j^{(x)}. \tag{3}$$

Here, $u_j^{(c)}$ and $u_j^{(x)}$ are cluster-level error terms or 'random effects', with the former often referred to as a 'random intercept' and the latter as a 'random slope' term. It is natural to think of these terms as capturing the effects of unmodelled cluster-level variables on $\beta_j^{(c)}$ and $\beta_j^{(x)}$. Typically, $u_j^{(c)}$ and $u_j^{(x)}$ are assumed to follow a multivariate normal distribution. Equation 2 is sometimes referred to as an 'intercept-as-outcome' equation and Equation 3 as a 'slope-as-outcome' equation.

Equations 1–3 highlight the multilevel nature of the model. An alternative formulation can be obtained by substituting Equations 2 and 3 into Equation 1. After rearranging terms we end up with:

$$Y_{ij} = \underbrace{\gamma^{(c)} + \gamma^{(cz)} z_j + \gamma^{(x)} x_{ij} + \gamma^{(xz)} z_j x_{ij}}_{\text{fixed part}} + \underbrace{u_j^{(c)} + u_j^{(x)} x_{ij} + \epsilon_{ij}}_{\text{random part } (=v_{ij})}. \tag{4}$$

Equation 4 shows why $\gamma^{(xz)}$ is referred to as a 'cross-level interaction effect': it is the coefficient on a multiplicative interaction term between the lower-level predictor x_{ij} and the cluster-level predictor z_j; in our running example, it is the interaction between the individual characteristic education and the country attribute HDI. The first part of the right-hand expression, consisting of the linear combination of the constant and the lower- and upper-level predictors, multiplied by their respective coefficients (or 'fixed effects'), is also referred to as the fixed part of the model. Crucially, the second part shows that the model has a complex error term v_{ij} that consists of three components: the random intercept term $u_j^{(c)}$, the lower-level residual error ϵ_{ij}, and the product of the random slope term with the lower-level predictor $u_j^{(x)} x_{ij}$.

Why *Always* a Random Slope?

The formal exposition of the multilevel model in the previous section provides an intuitive reason why one should *always* include the random slope term $u_j^{(x)}$: Equation 3 clarifies that omitting $u_j^{(x)}$ is equivalent to assuming that $\beta_j^{(x)}$ is perfectly determined by z_j, in other words that $R^2(\beta_j^{(x)})$, the R^2 of the (implicit) cluster-level regression for $\beta_j^{(x)}$, equals 1. As noted above, Raudenbush and Bryk (2002) do indeed discuss the possibility that 'little or no variance in the slopes remains to be explained' (p. 28) after accounting for the cluster-level predictor z_j. Yet we would argue that this is an unlikely scenario in the vast majority of social science applications. This is confirmed by the empirical examples presented in the 'Illustrative Empirical Analyses' section and in the Online Supplement (see, in particular, the final columns of Online Supplement Tables D1–D6). More importantly, our Monte Carlo simulations will show that omitting the random slope term can have severe consequences even when there is very little unexplained variation in $\beta_j^{(x)}$. We find that inference can be substantially overoptimistic even when $R^2(\beta_j^{(x)})$ is as high as 0.95 or when standard model selection criteria such as likelihood ratio tests or information criteria indicate that the remaining variation is negligible and favour the model that drops the random slope (the results on model selection strategies can be found in Online Supplement C).

The two-stage formulation of the model in Equations 1–3 also suggests that omission of $u_j^{(x)}$ should primarily affect inference about $\gamma^{(x)}$ and $\gamma^{(xz)}$ because these terms are implicitly defined in the potentially misspecified Equation 3. Statistical inference for estimates of $\gamma^{(cz)}$ and $\gamma^{(c)}$ should remain unaffected—as it should for any other terms that do not appear in Equation 3, including the coefficients of additional lower-level predictors.

We now further clarify the importance of including random slope terms on the lower-level components of cross-level interactions. Equation 4 shows that the presence of the random slope term $u_j^{(x)}$ in the true data-generating process (DGP) adds the component $u_j^{(x)} x_{ij}$ to the complex error term. This component has important consequences for the conditional variance of the overall error v_{ij} and for the covariance of the error terms for lower-level observations belonging to the same cluster. In particular, the variance of v_{ij} given x_{ij} will be (Snijders and Bosker, 2012, Equation 5.5):[1]

$$\begin{aligned}\mathrm{Var}(v_{ij}|x_{ij}) = {} & \mathrm{Var}(u_j^{(c)}) + 2\mathrm{Cov}(u_j^{(c)}, u_j^{(x)}) x_{ij} \\ & + \mathrm{Var}(u_j^{(x)}) x_{ij}^2 + \mathrm{Var}(\epsilon_{ij}).\end{aligned} \tag{5}$$

The covariance of the error terms for two different individuals (say, i and i') belonging to the same cluster will be (Snijders and Bosker, 2012, Equation 5.6):

$$\begin{aligned}\mathrm{Cov}(v_{ij}, v_{i'j}|x_{ij}, x_{i'j}) = {} & \mathrm{Var}(u_j^{(c)}) \\ & + \mathrm{Cov}(u_j^{(c)}, u_j^{(x)})(x_{ij} + x_{i'j}) \\ & + \mathrm{Var}(u_j^{(x)}) x_{ij} x_{i'j}.\end{aligned} \tag{6}$$

These equations highlight that v_{ij} will be heteroskedastic even if $u_j^{(c)}$, $u_j^{(x)}$ and ϵ_{ij} are all homoskedastic and that errors will be correlated within clusters. More specifically, if the true model includes the random slope term $u_j^{(x)}$, but the estimated model does not, there will be (a) unmodelled heteroskedasticity in the error term (due to the second and third term on the right-hand side in Equation 5) and (b) unmodelled covariation among the errors for lower-level observations belonging to the same cluster (due to the second and third term on the right-hand side in Equation 6).

Figure 1 illustrates the problem graphically. To construct the figure, we first simulated a data set according to Equations 1–3, assuming substantial cross-cluster variation in the slope of x_{ij}. We set the number of clusters to 25 and the number of lower-level observations per cluster to 100 (see the notes to Figure 1 for further information on how the data were generated). We then fitted a multilevel model with and a multilevel model without a random slope on x_{ij} to the simulated data and obtained the lower-level residuals for each. The figure plots these residuals against x_{ij} and $z_j x_{ij}$, after partialling out the cluster-level predictor z_j. We focus on three representative clusters, one with a slope for $\beta_j^{(x)}$ that deviates strongly positively from the average slope, one with a slope for $\beta_j^{(x)}$ that is close to the average (i.e., near-

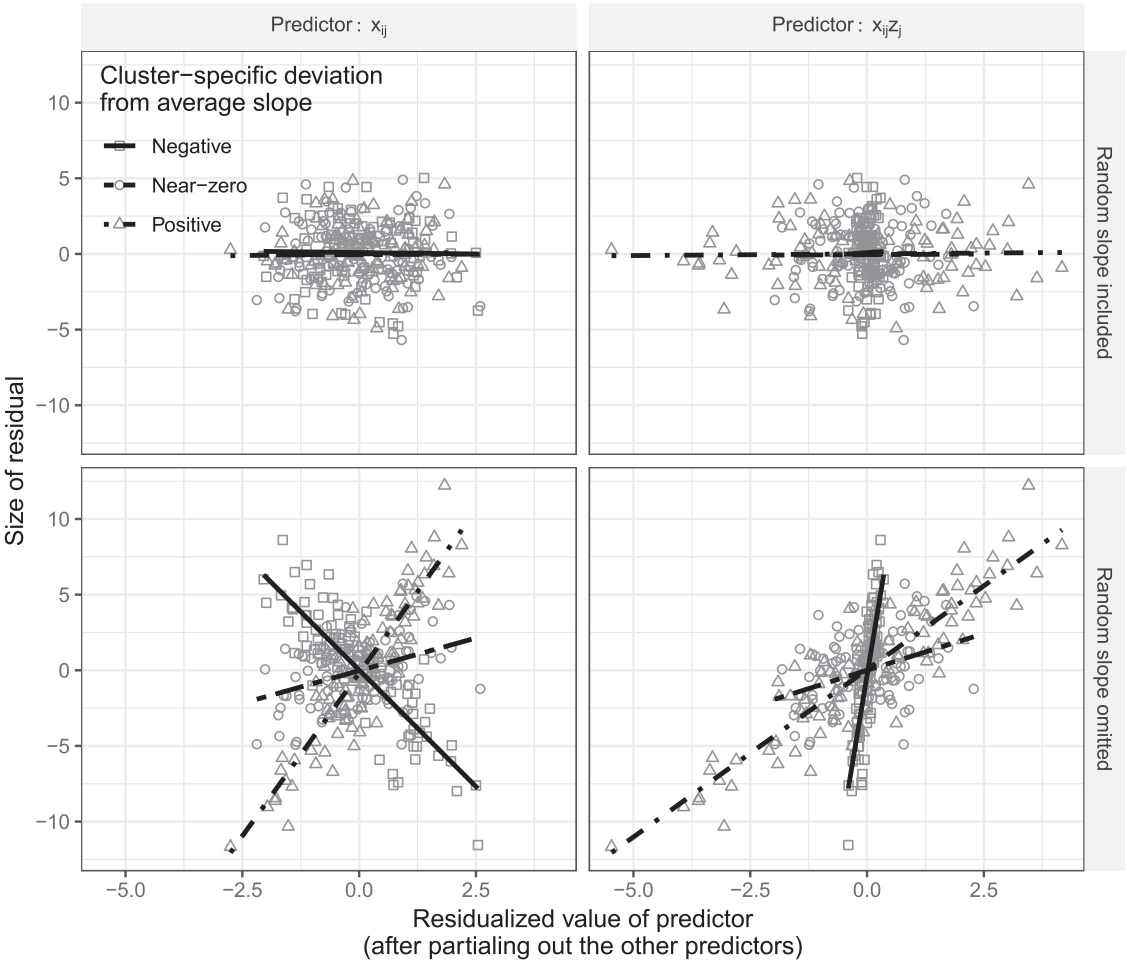

Figure 1. Lower-level residuals for models with and without random slope.

Notes: Residuals are from linear mixed-effects models. The data are simulated according to Equations 1–3 with 25 clusters and 100 lower-level observations per cluster. The cluster- and lower-level predictors, z_j and x_{ij}, are both normally distributed with means of 0 and standard deviations of 1 and their coefficients are being set to 1; $u_j^{(c)}$ and $u_j^{(x)}$ are multivariate normal with means of 0, standard deviations of 0.6 and 2, respectively, and with a correlation of 0.3; the lower-level error ϵ_{ij} is normally distributed with a standard deviation of 2.

zero), and one with a slope for $\beta_j^{(x)}$ that deviates strongly negatively from the average slope. Regression lines have been added to approximate the conditional mean of the residuals for each of the three clusters.

The graphs in the upper row of Figure 1 show that the lower-level residuals from the correctly specified model conform to the assumptions of the model: the cluster-specific means of the residuals are unrelated to either predictor and their variance is constant. The picture looks very different for the residuals from the misspecified model (i.e., the one omitting the random slope) in the bottom row. Consistent with the above discussion, the variance of the residuals is markedly higher for extreme values of x_{ij} (heteroskedasticity). Moreover, the residuals for lower-level observations belonging to the same cluster are highly positively correlated when they have similar values on x_{ij} and $z_j x_{ij}$.

Omission of a random slope that actually belongs in the model thus leads to unmodelled heteroskedasticity and unmodelled dependencies among the errors of units belonging to the same cluster. This will typically lead to the underestimation of standard errors and thereby to anti-conservative inference. This is well known not only from the multilevel modelling literature but also from the literature on cluster-robust inference in econometrics (for a recent overview, see Cameron and Miller, 2015).[2] In fact, the goal to achieve accurate inference in the presence of cluster-induced heteroskedasticity and cluster-correlated errors is a common motivation for both multilevel modelling and cluster-robust methods. The former approach seeks to address the interdependencies among observations belonging to the same cluster through the inclusion of random intercept and slope terms (see Equations 1–6 above). The latter uses special 'sandwich-type' estimators of the coefficient covariance matrix that remain consistent even in the presence of heteroskedasticity and cluster-correlated errors.

When will omitting the random slope term be particularly consequential? Inspection of Equations 5 and 6 (as well as Figure 1) suggests two relevant factors. First, the consequences of omitting the random slope should become more severe as the variance of $u_j^{(x)}$ increases. This is because both the conditional variance (Equation 5) and the within-cluster covariance (Equation 6) depend on $\mathrm{Var}(u_j^{(x)})$. The second factor is the extent of variation in the lower-level predictor, that is, $\mathrm{Var}(x_j^{(x)})$. As $\mathrm{Var}(x_j^{(x)})$ increases, so will the extent of (unmodelled) variation in the conditional error variance across observations. In terms of our running example, failure to model cross-cluster differences in the coefficient of education will be more consequential when individuals differ a lot in terms of their level of education.

The parallels to the literature on cluster-robust inference suggest a third factor that does not immediately follow from the above equations. The consequences of erroneously omitting the random slope term should also be related to the number of observations per cluster, that is, to the (average) cluster size. For the case of linear regression with clustered data, it is well known that the conventional (uncorrected) ordinary least squares variance estimate for a regressor x understates the true variance approximately by a factor of (Cameron and Miller, 2015: p. 322):

$$\tau \simeq 1 + \rho^{(x)} \rho^{(u)} (\bar{N}_g - 1), \tag{7}$$

where $\rho^{(x)}$ is the within-cluster correlation of x, $\rho^{(u)}$ is the within-cluster error correlation, and $\bar{N}_g$ is the average cluster size. Intuitively, the underlying reason is that the actual number of cases available for estimating the cross-level interaction is the number of clusters because the cross-level interaction is about a cluster-level relationship. This is immediately clear from the 'slope-as-outcome' formulation of the model (see Equation 3 above). By omitting the random slope term, this cluster-level nature of the cross-level interaction is ignored and observations from the same cluster are treated as contributing independent information about the moderating effect of z_j on the slope of x_{ij}. This illusionary increase in the number of cases available for estimating the cross-level interaction is larger when clusters are large.

In summary, the above discussion suggests that practitioners should always specify a random slope for the lower-level variable of a cross-level interaction in mixed-effects models. Failure to include a random slope is to disregard cluster-driven heteroskedasticity and within-cluster correlation among the errors, violating fundamental model assumptions. Omitting the random slope term associated with a cross-level interaction will not, in general, introduce systematic bias into coefficient estimates,[3] but it will lead to overly optimistic statistical inference for the cross-level interaction term and the coefficient (i.e., the 'main effect') of the lower-level variable involved in the interaction. All other coefficient estimates and their standard errors, including the main effect of the contextual predictor involved in the cross-level interaction as well as any additional lower- and upper-level predictors, should largely remain unaffected.[4] The consequences of omitting the random slope term should become more severe (a) as the unaccounted variation in the cluster-specific slopes grows, (b) as the variance of the involved lower-level variable increases, and (c) as the average cluster size becomes larger.

Inference for 'Pure' Lower-Level Effects

Against the background of the preceding discussion, one may wonder if the incorporation of random slopes is also important for achieving correct inference on the coefficients of lower-level variables that are not involved in a cross-level interaction term, that is, on 'pure' lower-level effects. In terms of our running example, this means: Does it remain crucial to include the random slope if we are interested in the overall (average) effect of education on fear of crime rather than the interaction between human development and education?

The likely answer to this question is *yes*. After all, it is the presence of an unmodelled random slope term $u_j^{(x)}$—and not the interaction between a cluster-level and a lower-level predictor—that introduces heteroskedasticity (Equation 5) and cluster-correlated errors (Equation 6) into the overall error term v_{ij}. To foreshadow our results, we do indeed find that the omission of a relevant random slope leads to anti-conservative inference also for pure lower-level effects. This is consistent with a recent study by Bell, Fairbrother and Jones (2018), who reach very similar conclusions concerning the case of pure lower-level effects but do not consider the case of cross-level interactions.[5]

This being said, we maintain and demonstrate below that there are at least two important reasons why the cross-level interaction case deserves special attention. The first is that, at least in sociology, the overwhelming majority of studies that use mixed-effects models with multilevel data are primarily interested in context effects, including cross-level interactions. The second reason is that the erroneous omission of a random slope term tends to be less consequential in the pure lower-level effect than in the cross-level interaction case. The reason for this is that, compared with a pure lower-level effect, much more data will usually be needed to achieve the same level of statistical power for identifying a cross-level interaction (Gelman and Hill, 2007: Ch. 20). As a consequence, the same relative increase in the standard error (due to omitting a random slope term) will often make the difference between moderately strong and no meaningful evidence against the null hypothesis in the cross-level interaction case (say, between $P < \cdot 0.05$ and $P > 0.1$). In the case of pure lower-level effects, the difference is more likely to be between different degrees of strong evidence (say, between $P < 0.001$ and $P < 0.01$). We further explore these issues in the 'Random Slopes and 'Pure' Lower-Level Effects' section, but in a first step we now turn to the Monte Carlo results for the cross-level interaction case.

Simulation Evidence

Simulation Set-Up

We now present Monte Carlo simulations to illustrate the importance of including random slopes alongside cross-level interaction terms. In Monte Carlo analysis, the statistical properties of competing estimators are evaluated under controlled conditions by repeatedly sampling data from a known DGP and applying the estimators to each simulated data set. By modifying key aspects of the DGP (e.g., the number of clusters), one can investigate how they shape the relative performance of the competing estimators.

The general form of the DGP for the simulations is given in Equations 1, 2, and 3 above. That is, we consider a simple case with one lower-level predictor x_{ij} and one upper-level predictor z_j, with the latter affecting both the constant and the slope of x_{ij}. In our running example, x_{ij} would be education, z_j would be human develompment, and the dependent variably y_{ij} would be fear of crime. We examine several variants of this DGP which, in keeping with standard terminology, we also refer to as 'experimental conditions'. In particular, we vary the number of clusters, the number of (lower-level) observations per cluster, the standard deviation of $u_j^{(x)}$ (the random slope term in Equation 3), and the extent of variability in the lower-level predictor x_{ij}. Table 1 lists the dimensions that we manipulate, along with the different values that we consider. In total, we analyse 162 ($= 3 \times 3 \times 6 \times 3$) experimental conditions. The coefficients on all predictors (i.e., $\gamma^{(cz)}$, $\gamma^{(x)}$, and $\gamma^{(xz)}$) are set to 1 and the overall constant $\gamma^{(c)}$ is set to 0. All predictors and random effects are normally distributed with means of 0. For their standard deviations, please see Table 1 and the replication files in the online supporting material.

We obtain 10,000 replications (i.e., 10,000 simulated data sets) per experimental condition and fit two mixed-effects models to each simulated data set. Consistent with the DGP, both models include the cluster-level predictor z_j, the lower-level predictor x_{ij}, and their cross-level interaction $z_j x_{ij}$. Both also include a random intercept term corresponding to $u_j^{(c)}$ in Equation 2. The only difference between the two models is that the first further includes a random slope term corresponding to $u_j^{(c)}$ in Equation 3, whereas the second model does not. As noted above, somewhat more than half of all cross-level interaction estimates published in the *ESR* between 2011 and 2016 are based on models that omit this random slope term (see also the 'Cross-Level Interactions in the *ESR*' section below).

Table 1. Dimensions manipulated in the Monte Carlo experiments

Dimension		Levels	$R^2(\beta_j^{(x)})$
Number of clusters	m	5	
		15	
		25	
Number of observations per cluster	n_g	100	
		500	
		1,000	
Standard deviation of random slope term $u_j^{(x)}$	$\mathrm{SD}(u_j^{(x)})$	0.1005	0.99
		0.1429	0.98
		0.2294	0.95
		0.3333	0.90
		1.0000	0.50
		3.0000	0.10
Standard deviation of lower-level predictor x_{ij}	$\mathrm{SD}(x_{ij})$	0.50	
		1.00	
		2.00	

Note: $R^2(\beta_j^{(x)})$ is the implied proportion of the overall cross-cluster variation in $\beta_j^{(x)}$ (the coefficient of the lower-level predictor x_{ij}) that is explained by the cluster-level predictor z_j.

We focus on statistical inference. There is no reason to expect that the omission versus inclusion of the random slope term affects parameter bias.[6] To assess inferential accuracy, we examine the actual coverage rates of two-sided 95 per cent confidence intervals. Accurate inference (for an unbiased estimator) requires that the actual coverage rate equals the nominal rate. We therefore examine whether two-sided 95 per cent confidence intervals cover the true parameter in more or less than 95 per cent of the 10,000 Monte Carlo replications. Let $C_{95}(r) = 1$ if the two-sided 95 per cent confidence interval for the r^{th} replication includes the true value of the parameter of interest and 0 otherwise. Then coverage is defined as

$$\text{Coverage} = \frac{1}{R}\sum_{r=1}^{R} C_{95}(r),$$

where R denotes the total number of replications. If coverage is greater than 95 per cent, confidence intervals are too large and over-conservative; hypothesis tests will retain the null hypothesis of no effect too often. By contrast, if coverage is below 95 per cent, confidence intervals are too narrow and null hypotheses rejected too frequently.

An alternative to the actual coverage rate would be to compare the average estimated standard error with the actual standard deviation of the corresponding point estimates across the Monte Carlo replicates (see, e.g., Schmidt-Catran and Fairbrother, 2016, who refer to this as 'optimism of the standard errors'). The reason why we prefer to measure accuracy in terms of the coverage rate is that the standard error is a (downward) biased estimator of the sampling distribution standard deviation in small samples. Since the work of William Gossett (i.e., Student, 1908), the established way of correcting for this downward bias is to base confidence intervals and hypothesis tests on an appropriate t-distribution rather than the standard normal distribution (as detailed below, we use the $m - l - 1$ rule advocated by Elff *et al.*, forthcoming, to select the appropriate t-distribution). We further explore these issues and present results on standard error optimism in Online Supplement A.

In practice, Monte Carlo estimates of actual coverage rates will typically differ from the ideal value even for accurate estimators because we use a finite number of Monte Carlo replications. In our case, 10,000 replications imply a simulation error of $\approx$ .00218 $(= \sqrt{0.95 \times 0.05/10000})$ or .218 percentage points. Thus, the actual coverage rate of an estimator (for a given experimental condition) is significantly different (at the five per cent level) from the nominal level of 95 per cent if it deviates from that level by more than 0.427 $(= 1.96 \times .218)$ percentage points. The null tested here is the hypothesis that the actual coverage rate is equal to the nominal rate.

We conducted all simulations in *R* (R Core Team, 2017), using the *lmer* function of the *lme4* package (Bates *et al.*, 2017) to estimate the mixed-effects models. Following the recommendations of Elff *et al.* (forthcoming), we use restricted maximum likelihood estimation throughout and construct confidence intervals based on a t-distribution with $m - l - 1$ degrees of freedom (where m represents the number of clusters and l generally equals 1 because we have only one cluster-level predictor). Replication files are available as part of the online supporting material.

Simulation Results

Table 2 shows actual coverage rates for models that omit versus models that include a random slope term on the lower-level component of the cross-level interaction. Results are displayed along two dimensions: the amount of unexplained variation in the cluster-specific slope $\beta_j^{(x)}$ and the extent of variation in x_{ij}. The number of clusters is 15 and the number of lower-level observations per cluster is 500 throughout the table. We explore the impact of varying these factors below.

The central result in Table 2 is that coverage rates of confidence intervals based on models that omit the random slope term are inaccurate. As expected, this does

Table 2. Actual coverage rates of nominal 95 per cent confidence interval by variance of lower-level predictor and random slope term

SD(x_{ij})	$\gamma^{(x)}$ Random slope		$\gamma^{(xz})$ Random slope		$\gamma^{(cz})$ Random slope	
	Included	**Omitted**	**Included**	**Omitted**	**Included**	**Omitted**
			$R^2(\beta_j^{(x)}) = 0.95$ (i.e., SD$(u_j^{(x)}) \approx 0.23$)			
0.5	96.44	92.82	96.46	93.07	95.17	95.21
1.0	95.21	81.74	95.12	81.69	94.79	95.07
2.0	95.00	57.38	94.60	56.95	94.85	95.20
			$R^2(\beta_j^{(x)}) = 0.90$ (i.e., SD$(u_j^{(x)}) \approx 0.33$)			
0.5	95.54	88.55	95.64	88.33	94.74	94.75
1.0	95.34	70.06	95.12	68.55	94.89	95.20
2.0	95.04	42.11	95.23	43.34	94.51	95.03
			$R^2(\beta_j^{(x)}) = 0.50$ (i.e., SD$(u_j^{(x)}) = 1.00$)			
0.5	95.14	53.94	95.32	54.28	94.74	95.10
1.0	94.90	30.55	94.84	30.51	94.80	95.19
2.0	95.00	17.14	94.74	17.15	94.95	95.03
			$R^2(\beta_j^{(x)}) = 0.10$ (i.e., SD$(u_j^{(x)}) = 3.00$)			
0.5	94.74	21.87	94.84	21.16	94.95	94.82
1.0	95.03	12.54	95.12	12.87	95.20	95.13
2.0	94.78	8.85	95.21	8.78	94.98	95.38

Notes: Results are based on 10,000 Monte Carlo replications. Because of Monte Carlo sampling error, the 95 per cent test interval is 95±0.427. Values smaller or larger than that are statistically significantly different (five per cent level) from the nominal coverage rate of 95 per cent. The number of observations per cluster is 500 with overall 15 clusters.

not apply to inference for the main effect of the contextual predictor z_j where coverage rates fall within the range of 95±0.427 per cent for all experimental conditions. But the coverage rates of confidence intervals for the cross-level interaction term and for the main effect of the lower-level predictor are too low and the extent of undercoverage is generally substantial. To understand the implications, note that an actual coverage rate of 90 per cent means that nominal significance on the 5 per cent level would actually only mean 'marginal' significance on the 10 per cent level.[7] Yet, most actual coverage rates displayed in Table 2 are even substantially smaller than 90 per cent. Our simulation results therefore suggest that omitting the random slope term can easily turn coefficient estimates that are actually far from any conventional level of statistical significance into ones that seemingly surpass the corresponding thresholds. Results for standard error optimism in Online Supplement A are qualitatively similar. Estimated standard errors for the cross-level interaction term and the main effect of the lower-level variable exhibit substantial downward bias when the model does not include the random slope term: The estimated standard errors are systematically smaller than the true standard deviation of the corresponding point estimates, meaning that they overstate the precision with which these coefficients can be estimated.

By contrast, coverage rates of confidence intervals based on models that include the random slope term are by and large accurate for all three coefficients and across all displayed experimental conditions. Only when variation is low for both the lower-level predictor (i.e., SD(x_{ij})) and the random slope term (i.e., SD$(u_j^{(x)})$) do the results show a tendency for overly conservative inference, meaning that confidence intervals might be somewhat too wide (see, in particular, the results for the case where SD$(x_{ij}) = 0.5$ and SD$(u_j^{(x)}) = 0.23$ in the first row of Table 2). We return to this unexpected result at the end of this section.

The next important question is: What drives the extent of miscoverage? As expected, the extent of undercoverage grows with the unaccounted cluster-specific variation of $\beta_j^{(x)}$ in the true model (i.e., with SD$(u_j^{(x)})$) and also with the extent of variation in x_{ij} (i.e., with SD(x_{ij})). Equations 5 and 6 above show why: The extent of heteroskedasticity and within-cluster error correlation that remains unmodelled in the specification that omits the random slope is a function of the product of these two factors (i.e., of SD$(u_j^{(x)})$ and SD(x_{ij})). This is also why each dimension on its own can drive the extent of undercoverage to completely unacceptable levels.

We further argued that the (average) size of the upper-level units or ‘clusters’ should exacerbate the consequences of omitting a random slope term because models without a random slope term assume too much independence among observations (see discussion of Equation 7 above). We explore this issue in Table 3, which shows actual coverage rates by the number of clusters and number of observations per cluster. SD(x_{ij}) is set to 1 and the implicit cluster-level $R^2(\beta_j^{(x)})$ to 0.5 (i.e., SD($u_j^{(x)}$) = 1.00); that is, we hold both factors at the intermediate levels considered in Table 2 above.

Table 3 confirms that inference based on models that include a random slope is generally accurate, although we find some very limited deviations from the ideal value of 95 per cent when n_j, the number of observations per cluster, equals 100. As before, we also see that omitting the random slope term does not, in general, compromise inference for $\gamma^{(cz)}$, while confidence intervals for $\gamma^{(xz)}$ and $\gamma^{(xz)}$ exhibit substantial undercoverage. As expected, the problem gets worse as the cluster size (i.e., n_j) increases. For every given number of clusters, undercoverage is most severe for 1,000 observations per cluster as compared with 500 and especially 100 observations per cluster.

The upshot of our Monte Carlo simulations thus is that omitting the random slope term on the lower-level component of a cross-level interaction can lead to dramatically anti-conservative statistical inference for the interaction term and the main effect of the lower-level variable. In line with our expectations, undercoverage increases with the extent of variation in the lower-level variable, the extent of variation in the unmodelled random slope term, and the (average) size of the clusters.

Before we investigate the severity of the problem using real-life data from the ESS, we summarize the main results of two additional sets of simulations.

In Online Supplement B, we further investigate the unexpected result that the (correctly specified) model including the random slope term yields over-conservative statistical inference in some situations. We present additional simulations that consider even lower values of 0.14 and 0.10 for the standard deviation of the random slope term, implying values of 0.98 and 0.99 for the cluster-level $R^2(\beta_j^{(x)})$. The additional simulations confirm that very low variation in the random slope term can lead to substantial overcoverage, especially when the number of clusters is also very low. While these results do warrant a note of caution, their practical relevance is limited. In the vast majority of applications, the number of clusters is at least in the tens, and cross-cluster variation in random slopes is typically substantial, at least in country-comparative setting. This is confirmed by the empirical examples presented in the next section and in the Online Supplement (see, in particular, the final columns of Online Supplement Tables D1–D6). Moreover, practitioners can easily verify if they are dealing with a situation where the random slope variation is close to 0.

Table 3. Actual coverage rates (per cent) of nominal 95 per cent confidence interval by number of clusters and lower-level observations

		$\gamma^{(x)}$ Random slope		$\gamma^{(xz)}$) Random slope		$\gamma^{(cz)}$) Random slope	
n_j	n_{total}	Included	Omitted	Included	Omitted	Included	Omitted
			$m = 5$ Clusters				
100	500	96.20	77.16	96.18	77.45	97.35	97.82
500	2,500	95.09	43.23	95.07	43.68	93.64	95.34
1,000	5,000	95.07	31.39	94.58	31.70	93.95	95.11
			$m = 15$ Clusters				
100	1,500	95.19	58.57	94.75	58.62	93.65	95.51
500	7,500	94.90	30.55	94.84	30.51	94.80	95.19
1,000	15,000	94.93	21.46	94.95	22.37	95.10	95.25
			$m = 25$ Clusters				
100	2,500	94.79	56.87	95.24	56.22	93.23	95.03
500	12,500	94.93	29.71	94.98	29.29	95.13	95.14
1,000	25,000	94.85	21.43	95.23	21.32	94.90	94.74

Notes: Results are based on 10,000 Monte Carlo replications. Because of Monte Carlo sampling error, the 95 per cent test interval is 95±0.427. Values smaller or larger than that are statistically significantly different (five per cent level) from the nominal coverage rate of 95 per cent. These results are based on experimental conditions for which $R^2(\beta_j^{(x)}) = 0.50$ (i.e., SD($u_j^{(x)}$) = 1) and SD(x_{ij}) = 1.

In a second set of supplementary analyses, presented in Online Supplement C, we investigate the performance of a data-driven approach to model selection. As noted in the introduction, Raudenbush and Bryk (2002: p. 28) suggest that it might be appropriate to omit the random slope if its variance is 'very close to zero'. For want of an exact definition of 'very close', one might turn to standard model selection criteria for determining whether a given slope is small enough to warrant omission. Our supplementary analyses consider four selection criteria: Akaike's information criterion, the Bayesian information criterion, and two variants of a likelihood ratio test. The main result is unambiguous: when the goal is to achieve correct statistical inference for a cross-level interaction effect, it is not advisable to rely on model selection criteria in deciding whether to include a random slope on the lower-level predictor. For all four selection criteria, we find settings where reliance on the criterion results in noteworthy levels of undercoverage.

Illustrative Empirical Analyses

The simulation results are clear cut: Omitting random slopes on the lower-level components of cross-level interaction terms compromises statistical inference about those terms and about the main effects of their lower-level components. To get a better sense of how serious the problem is in real-world applications, we now present a series of illustrative country-comparative analyses based on ESS data (ESS Round 6, 2016). As noted above, such (cross-sectional) country comparisons are by far the most common type of multilevel analysis published in the *ESR*.

We adopt Heisig, Schaeffer and Giesecke's (2017) illustrative analyses of cross-level interactions.[8] The overall 30 empirical examples study how the relationships between six lower-level predictors (having a high level of education, age, gender, unemployment, being married, and having an intermediate level of education) and five outcome variables (generalized trust, homophobia, xenophobia, fear of crime, and occupational status) are moderated by the HDI.

For each of the 30 cross-level interactions (5 dependent variables × 6 lower-level predictors), we estimate two specifications, resulting in a total of 60 linear mixed-effects models. The first specification is a random intercept and slope model that assigns a random effect to the coefficient of the lower-level variable involved in the focal cross-level interaction. According to our simulation evidence, this model is correctly specified. The second is a random intercept model without any random slopes. This model is widespread in applied research, but the above analysis shows that it is misspecified and provides anticonservative inference for the cross-level interaction term and the main effect of its lower-level component. In addition to the lower-level predictor of interest, the HDI, and their cross-level interaction, the models always contain the other lower-level predictors as control variables. Online Supplement D gives a brief description of the coding of the variables and provides exact results for the coefficients of interest in Online Supplement Tables D1–D6. For brevity, we focus on statistical inference for the cross-level interaction term in the main article. In line with the simulation evidence, results are similar for the main effect of the lower-level predictor, while omitting the random slope term has no consequences for statistical inference about the main effect of the upper-level moderator.

Figure 2 illustrates the main results. It shows, for each of the 30 cross-level interactions, by how much the absolute *t*-ratio changes when a random slope is included. Changes are shown as directed arrows on a logged scale, with the origin of the arrow denoting the *t*-statistic for the model omitting and the head denoting the *t*-statistic for the model including the random slope.

Nearly all arrows point downwards, indicating that absolute *t*-ratios for the models including the random slope term are lower, and often very substantially so. Take our running example, for instance, which is expressed by the second arrow from the right. The model which does not contain a random slope on high education yields an absolute *t*-ratio of 9.7 for the cross-level interaction between having high education and the HDI on fear of crime. The corresponding value for the model including the random slope is only 5.1, a reduction of 46.8 per cent (see Online Supplement D for these values and the associated point estimates). Figure 2 shows that reductions of such alarming magnitude are the rule rather than the exception (because the y-scale is logged arrows of similar length indicate similar *relative* changes). Over the 30 different models, the reduction in the absolute *t*-ratio for the cross-level interaction effect due to including the random slope is 42.4 per cent on average. The median reduction is 48.3 per cent and the 25th and 75th percentiles are 31.3 per cent and 60.9 per cent, respectively.

The final columns of Online Supplement Tables D1–D6 convey another important result. They display the remaining variation of the random slope in the model including the cross-level interaction, expressed as the ratio of the random slope standard deviation to the corresponding main effect. Thus, the values are directly comparable with the values of $\mathrm{SD}(u_j^{(x)})$ in our Monte Carlo simulations. Remaining variation in the random

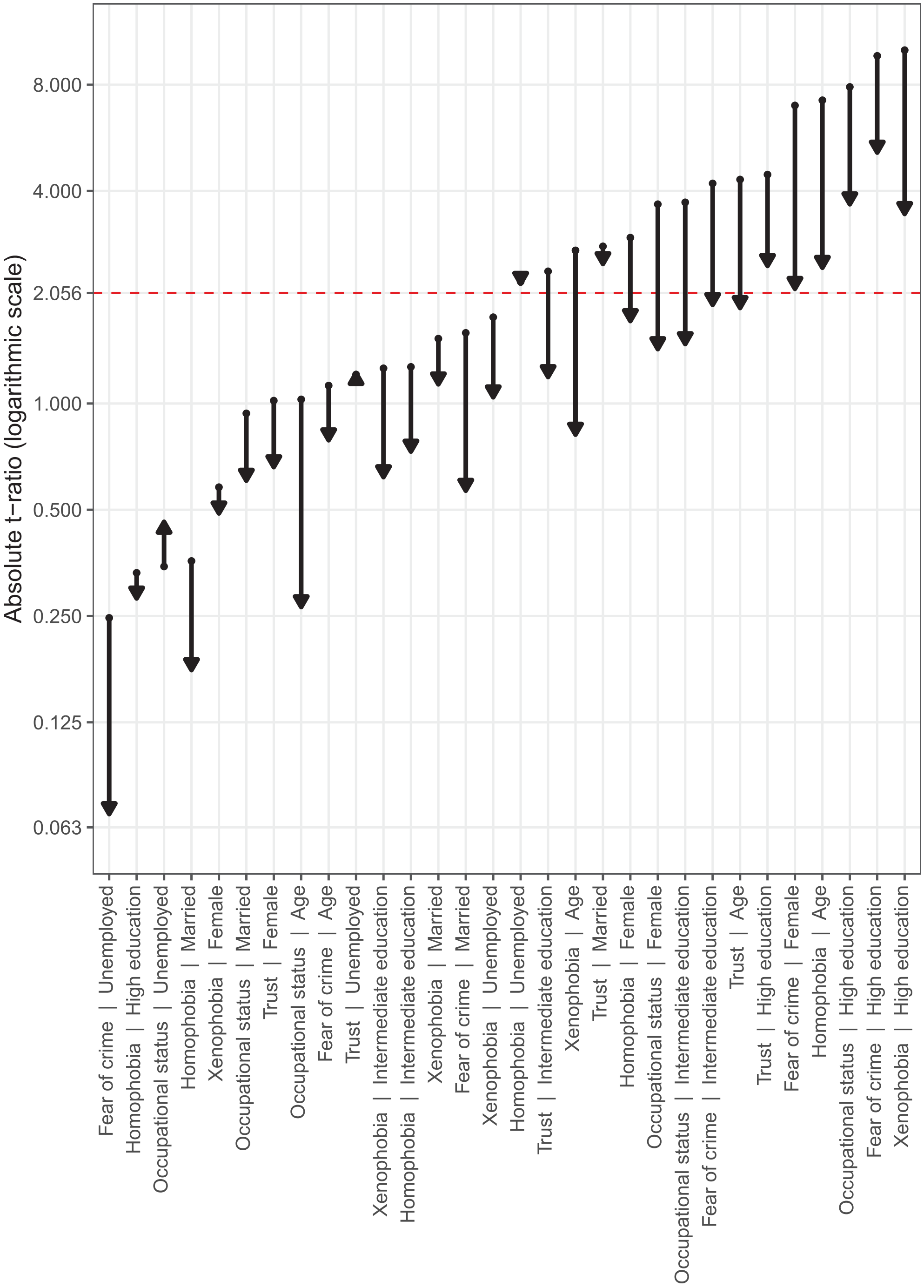

Figure 2. Changes in absolute *t*-ratios for 30 prototypical cross-level interactions after inclusion of random slopes expressed as directed arrows.

Notes: The triangled arrow heads show the absolute *t*-ratio from the specification including a random slope for the lower-level predictor of a cross-level interaction. The point start of the arrows indicates the absolute *t*-ratio from the specification omitting the random slope. The labels name the outcome (e.g., fear of crime) and lower-level predictor involved in the cross-level interaction (e.g., unemployed). The country-level moderator is always the HDI. The overall 60 cross-level interactions are estimated by linear mixed-effects models, which are displayed in Online Supplement Tables D1–D6. The dashed horizontal line demarcates 2.056, the threshold for statistical significance at the five per cent level (two-tailed test). The threshold is based on a *t*-distribution with 26 (=28 − 2) degrees of freedom, as suggested by Elff *et al.* (forthcoming).

slope term is substantial for most of our 30 illustrative analyses (mean = 2.03; median = 0.78, $p_{25} = 0.38$, $p_{75} = 1.61$). Hence, the model including the random slope is unlikely to suffer from overcoverage (see the discussion in the previous section and in Online Supplement B).

We have discussed the results of the empirical illustrations primarily in terms of changes in *t*-statistics and significance. However, we would like to emphasize that the inclusion versus omission of the random slope matters for the accurate assessment of statistical uncertainty more broadly. Even if the omission of the random slope term does not lead to a change in statistical significance, it will lead to standard errors that are too small and confidence intervals that are too narrow.

Against these results, we conclude that not specifying random slopes on the lower-level components leads to invalid statistical inference about cross-level interactions—and that the magnitude of the problem will be considerable in many sociological applications.

Cross-Level Interactions in the *ESR*

Given our findings, one may wonder whether current multilevel modelling practice meets the requirements for correct inference by including random slopes on the lower-level components of cross-level interactions. To answer this question, we reviewed all articles that investigate a cross-level interaction and that were published in the *ESR* between 2011 and 2016. We confined ourselves to studies using simple two-level models where lower-level observations are nested in exactly one type of upper-level unit. We identified 28 studies, the vast majority of which (24 or 86 per cent) were country comparisons (one of the remaining studies treated individuals as nested in combinations of countries and survey years). The 28 studies reported a total of 150 estimates of cross-level interactions. Some studies provided multiple estimates of the same cross-level interaction (i.e., of the same combination of lower-level, cluster-level, and outcome variable), for example, because they compared results across different subsamples or sets of control variables. We chose one estimate at random in these cases. For brevity, we continue to restrict our attention to cross-level interaction effects and do not consider the estimated main effects of the cluster- and lower-level components in this section because the cross-level interaction terms tend to be of primary interest to authors.

The discomforting result of our review is that not even half of the studies (11/28 or 39 per cent) specified random slopes on the lower-level components of the cross-level interactions they investigate. Figure 3 displays the percentage of studies that included random slope terms by year of publication. It provides no evidence that correct specifications have become more popular over time. As there is little reason to suspect that these problems are confined to articles that appeared in the *ESR*, we conclude that a large number of published sociological studies fail to meet the requirements for correct statistical inference about cross-level interactions.

We have shown that inclusion of random slopes on the lower-level components of cross-level interactions results in larger standard errors and smaller absolute *t*-ratios, so studies using the correct random-effects structure should be less likely to find statistically significant effects. To investigate this implication, we surveyed inferential statistics for the 150 cross-level interactions estimated in the 28 *ESR* articles. If available, we collected the *t*-ratio and otherwise the *P*-value or point estimate and standard error to compute the *t*-ratio from these statistics.[9] Unfortunately, several studies only report whether the estimated cross-level interactions attain a certain level of statistical significance, such as the 5 per cent level of significance, as commonly indicated by a single asterisk $*$.[10] Another problem is the rounding of point estimates and standard errors, especially in combination with many leading zeros, which often result in tiny coefficients and tiny standard errors which are then rounded and reported as ‘0.00’. In such extreme cases, it is impossible to reliably approximate the *t*-statistic and we again surveyed the level of significance of the cross-level interaction term.

Table 4 displays the percentage of estimated cross-level interaction effects that attain a given level of statistical significance according to whether the model did or did not include a random slope on the lower-level component. It shows a consistent pattern of more insignificant results for models that are correctly specified and include the random slope. By contrast, marginally significant and especially significant and highly significant results were less likely to occur when the random slope term was included. This is exactly what our arguments, Monte Carlo simulations, and illustrative empirical analyses would suggest. Nevertheless, the pattern appears less pronounced than one might expect given the results of our simulations and exemplary analyses. An important factor to consider in this regard is potential publication bias against insignificant findings, which obviously hits correctly specified cross-level interactions more often because their standard errors are not deflated. In other words, a larger share of correctly estimated cross-level interactions most likely never made it into the *ESR*, although proving this is difficult because

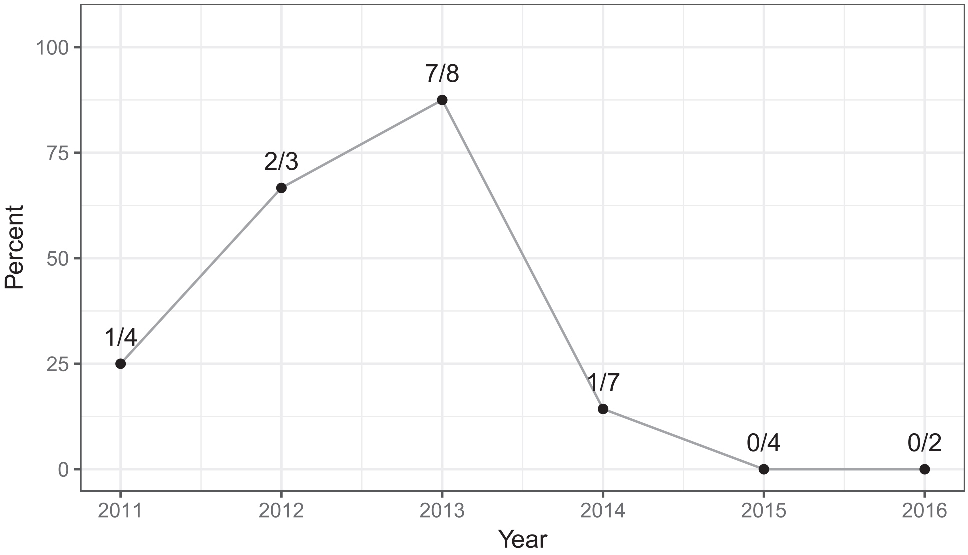

Figure 3. Proportion of articles that include a random slope on the lower-level components of cross-level interaction terms.

Note: Results are based on 28 articles reporting cross-level interaction terms from two-level mixed-effects models published in the *ESR*, 2011–2016.

Table 4. Percent of cross-level interaction terms by surpassed significance levels

		Random slope	
		Included	Omitted
Insignificant	($P \geq 0.1$)	64.71	42.42
Marginally significant	($P< 0.1$)	1.96	2.02
Significant	($P< 0.05$)	13.73	22.22
Highly significant	($P< 0.01$)	19.61	33.33
		100.00	100.00
Overall ($n = 150$)		($n = 51$)	($n = 99$)

Notes: Results are based on 28 articles reporting 150 cross-level interactions from two-level mixed-effects models published in the *ESR* 2011–2016. As many articles did not report levels of significance beyond $P < 0.01$, we restrict our review to this threshold as the highest level of significance.

about 60 per cent of null results are never written up (Franco, Malhotra and Simonovits, 2014). Online Supplement E uses *P*-curve analyses following Simonsohn, Nelson and Simmons (2014, 2015) to more systematically investigate the possibility of publication bias and '*P*-hacking' (i.e., selective reporting of subsets of analyses that yield significant results).

Another important question is how many findings should never have made it into the *ESR*, at least not as evidence of a statistically significant cross-level interaction?[11] We cannot give a definitive answer to this question based on published regression output—this would require actual reanalysis of the published studies. But in combination with our simulation evidence and the illustrative empirical analyses, Figure 4 allows us to make an informed speculation. The figure shows the distribution of absolute *t*-ratios for the 86 cross-level interaction terms where this information was provided or where we could at least obtain a good approximation. The upper panel shows the distribution of *t*-ratios from misspecified models that omit the crucial random slope term. The lower panel shows the distribution from models that include it.

Figure 4 shows a pronounced peak near the threshold for statistical significance at the 5 per cent level ($t = 1.96$). This unnatural peak characterizes the distribution of *t*-ratios especially for the incorrectly specified models and is suggestive of *P*-hacking. Online Supplement E further investigates this issue and finds some aggregate-level evidence for *P*-hacking among studies that did not specify random slopes for their cross-level interactions but not among those that correctly included a random slope.

What matters here more immediately is another implication of the clustering of *t*-ratios just above 1.96: In light of the above evidence, it seems almost certain that the line for cross-level interactions tested without a random slope needs to be shifted substantially to the left. That is, the true *t*-ratios for the cross-level interactions that were estimated using such models will often be much smaller. If we take the illustrative empirical

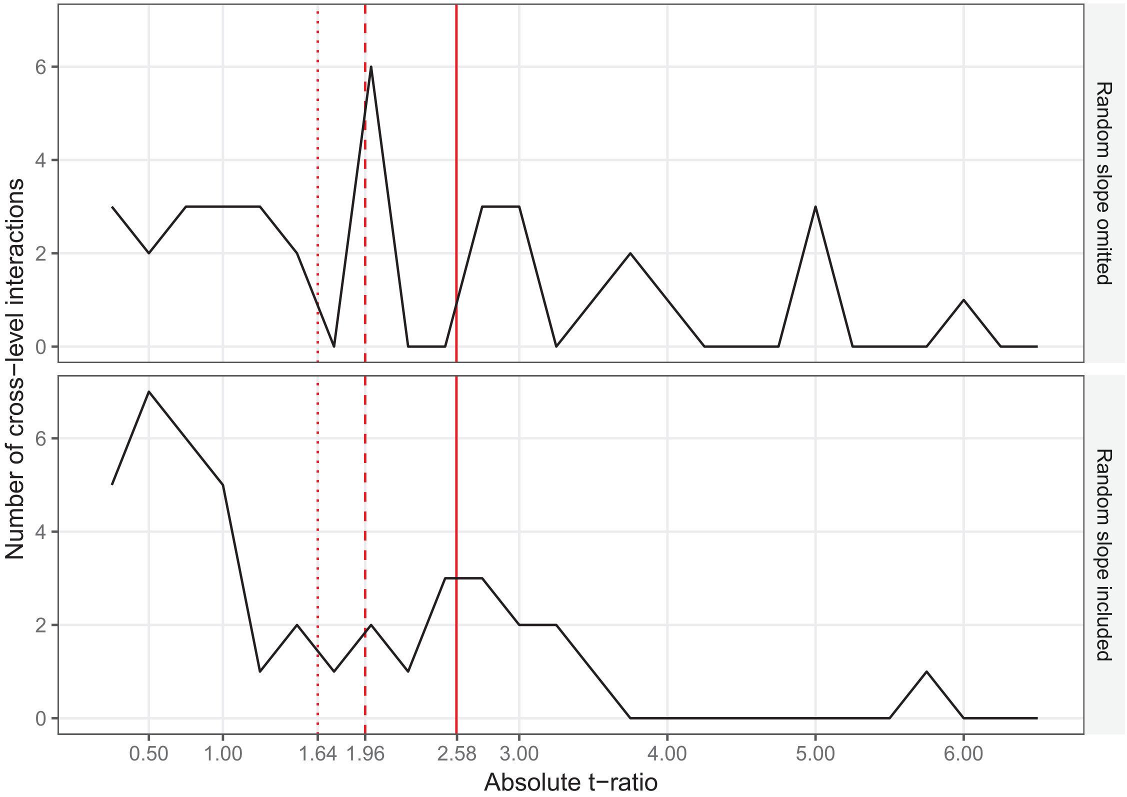

Figure 4. Distribution of absolute *t*-ratios of cross-level interactions.

Notes: Results are based on 86 cross-level interaction terms from two-level mixed-effects models reported in 20 articles that were published in the *ESR* between 2011 and 2016. The number of interaction terms and articles is lower than that in Table 4 because we could only include articles where authors reported *t*-statistics or for which we were able to approximate them reasonably well. Bin width is set to 0.25.

analyses at face value, the correct *t*-ratios will be at least 31 per cent smaller for three quarters of these estimates (cf. the percentiles of the relative reductions in *t*-ratios reported above). This suggests that many of the cross-level interaction effects based on misspecified models are not actually statistically significant at conventional levels. Thus, they should probably not have made it into the *ESR* or at least should have been interpreted very cautiously.

This conclusion is further reinforced if we take into account that critical values based on the normal distribution (i.e., $t = 1.96$ for $P < 0.05$ and $t = 2.58$ for $P < 0.01$) are questionable when cluster-level samples are small. Elff *et al.* (forthcoming) elaborate that critical values for cross-level interaction terms should instead be derived from a *t*-distribution with the appropriate degrees of freedom typically being smaller than the number of clusters. Given that many of the surveyed studies work with cluster-level sample sizes in the 10s or 20s, this recommendation would often result in substantially larger critical values. As this problem also applies to the cross-level interaction terms that were estimated including a random slope, one has to wonder how much robust evidence of cross-level interactions European sociology has generated at all.

Random Slopes and 'Pure' Lower-Level Effects

The results so far compellingly demonstrate that inclusion of a random slope term on the lower-level component is crucial for achieving correct statistical inference about the cross-level interaction term and the main effect of the lower-level variable. A natural follow-up question is whether the random slope term is also important for inference on the coefficients of lower-level variables that are not involved in a cross-level interaction, that is, for 'pure' lower-level effects. We showed above that omitting a random slope that is actually present in the DGP introduces heteroskedasticity (Equation 5) and within-cluster correlation (Equation 6) into the overall error term v_{ij}, and importantly, this fact does not hinge on the presence of a cross-level interaction term in the DGP.

Further Monte Carlo simulations indeed show that the inclusion of random slope terms is also essential for inference about pure lower-level effects. The basic DGP and the experimental conditions considered in these further analyses are identical to those presented in the 'Simulation Evidence' section above. There is only one crucial difference, namely, that $\beta_j^{(x)}$, the coefficient on the lower-level variable x_{ij}, no longer depends on the cluster-level predictor z_j (in other words, the DGP no longer includes a cross-level interaction):

$$\beta_j^{(x)} = \gamma^{(x)} + u_j^{(x)}. \tag{8}$$

Table 5 shows results for the same experimental conditions as Table 2. It yields virtually identical conclusions. When the coefficient of a lower-level variable varies across clusters, statistical inference for the (average) coefficient, i.e., for $\gamma^{(x)}$, will be anti-conservative unless that variation is captured by a random slope term. As in the cross-level interaction case, the problem becomes worse as the extent of cross-cluster variation in the lower-level effect increases (i.e., the higher SD($u_j^{(x)}$) is). Moreover, because the source of the problem is heteroskedasticity that correlates with x_{ij}, more variation in x_{ij} amplifies the inaccuracy of statistical inference with respect to $\gamma^{(x)}$. Online Supplement Table F1 further reaffirms that the average cluster size exacerbates the problem, just as in the cross-level interaction case (see Table 3 above). Across all experimental conditions, the extent of statistical over-confidence, as measured by the undercoverage of two-sided 95 per cent confidence intervals, is generally very similar to the corresponding results for the cross-level interaction case.

Table 5. Actual coverage rates of nominal 95 per cent confidence interval by variance of lower-level predictor and random slope term

SD(x_{ij})	$\gamma^{(x)}$	
	Random slope	
	Included	Omitted
	SD($u_j^{(x)}$) ≈ 0.23	
0.5	96.43	93.16
1.0	95.60	81.58
2.0	95.17	56.26
	SD($u_j^{(x)}$) ≈ 0.33	
0.5	95.50	88.82
1.0	95.47	69.53
2.0	94.79	41.94
	SD($u_j^{(x)}$) = 1.00	
0.5	95.17	53.88
1.0	94.89	30.52
2.0	95.01	17.19
	SD($u_j^{(x)}$) = 3.00	
0.5	95.23	21.18
1.0	95.13	12.29
2.0	95.20	8.55

Notes: Results are based on 10,000 Monte Carlo replications. Because of Monte Carlo sampling error, the 95 per cent test interval is 95±0.427. Values smaller or larger than that are statistically significant deviations and indicate biased inference. The number of observations per cluster is 500 with overall 15 clusters.

Despite these results, we maintain that the cross-level interaction case is more problematic and deserves special attention for at least two reasons. First, practitioners who analyse multilevel data with mixed-effects models are primarily interested in context effects. Second, lower-level effects tend to be so precisely estimated that inaccurate inference is less likely to lead to qualitatively different conclusions. We now elaborate on both of these issues.

Our reading of applied research using mixed-effects multilevel models is that practitioners predominantly use these models to test hypotheses about context effects. Typically, lower-level variables are mainly included to adjust for compositional differences among clusters. So while inference for lower-level effects might be over-confident, it rarely matters for the main research questions. To check the accuracy of this impression, we extended our review of *ESR* articles that used (two-level) mixed-effects models and were published between 2011 and 2016. For each article, we coded whether (i) the title, (ii) the abstract, and (if existent) (iii) explicitly formulated hypotheses stress (a) individual-level relationships, (b) contextual relationships (direct context effects and/or cross-level interactions), or (c) both.

Table 6 shows the results. The number of studies differs across the columns of the table because it was not always possible to classify a given article. For example, an article might not include any explicit hypotheses or the title of an article might mention neither lower-level nor contextual relationships. The first column of Table 6 indicates that only 3 out of 56 articles (5.4 per cent) using (two-level) mixed-effects models exclusively posit hypotheses about lower-level effects. By contrast, 53.6 per cent formulate hypothesis about both pure lower-level and contextual relationships and 41.1 per cent only present hypotheses about contextual

Table 6. Percent of articles testing context or lower-level effects

	Explicit hypotheses	Abstract	Title
Context effects	41.07	50.00	66.67
Lower-level effects	5.36	6.06	18.75
Both	53.57	43.94	14.58
n	56	66	48

Notes: Results are based on 68 articles using two-level mixed-effects models published in the *ESR* 2011–2016. Because of missing values (i.e., difficulties to decisively code), the numbers (*n*) of coded hypotheses, abstracts, and titles differ.

relationships. A similar pattern emerges if we consider the abstracts of the articles. In some sense, these figures may even overstate the salience of pure lower-level effects in the surveyed studies. Our impression from coding the articles is that hypotheses about lower-level relationships are often the ones that are least novel and that authors take the least interest in. This is also why, as we turn to titles, where authors are forced to stress the cardinal contribution of their article, the mixed category shrinks to ca. 15 per cent—mostly because articles tend to highlight only context effects in their title. Two thirds of all articles fall into this category.

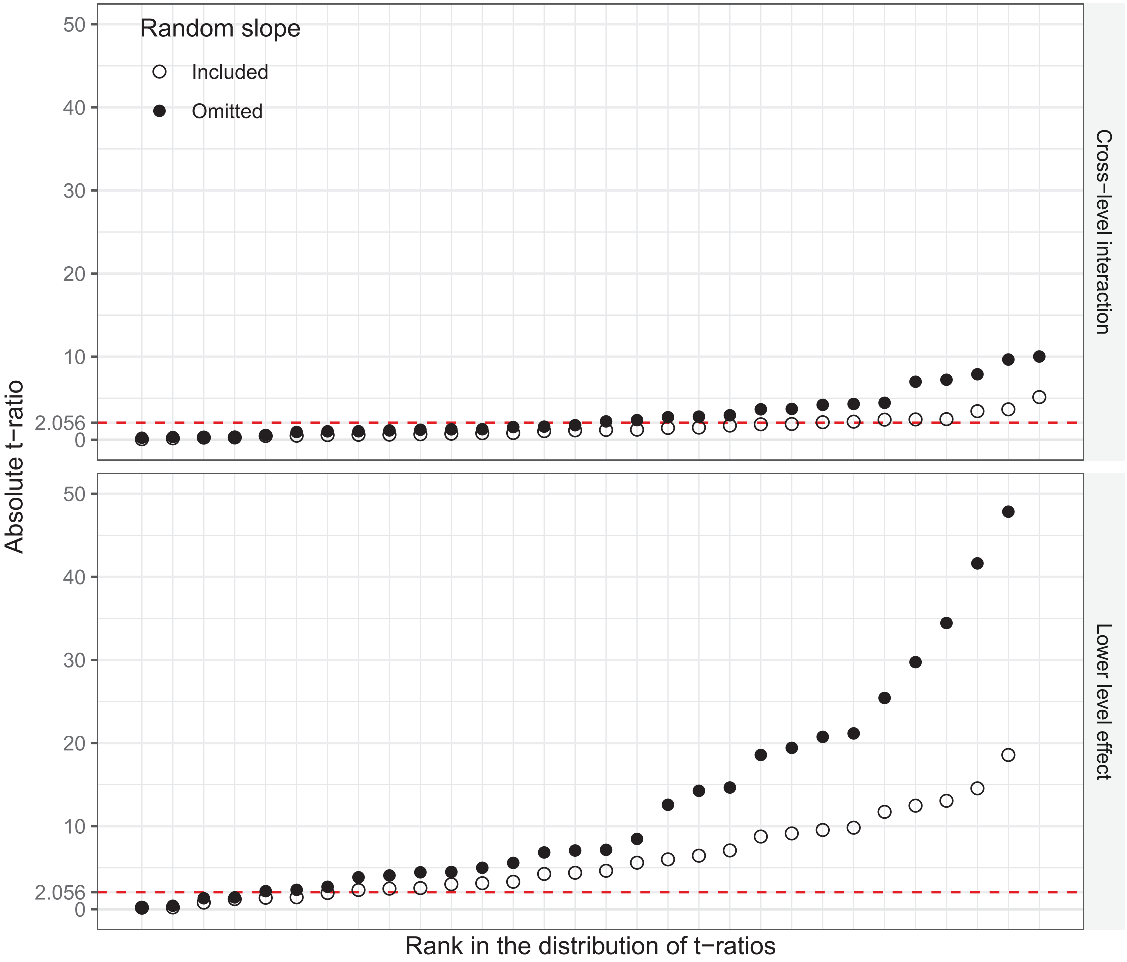

Figure 5. Distribution of absolute *t*-ratios.

Notes: The 60 absolute *t*-ratios for cross-level interactions are estimated by linear mixed-effects models, which are displayed in Online Supplement Tables D1–D6. The 60 absolute *t*-ratios for lower-level effects are estimated by identical linear mixed-effects models that simply omit the cross-level interaction terms. Note that 2 of the 60 *t*-ratios for pure lower-level effects are omitted from the bottom panel. The reason is that these two cases are extreme outliers with absolute *t*-ratios of approximately 142 for the model omitting and 66 for the model including the random slope. The dashed horizontal line demarcates 2.056, the threshold for statistical significance at the five per cent level (two-tailed test). The threshold is based on a *t*-distribution with 26 (=28 − 2) degrees of freedom, as suggested by Elff *et al.* (forthcoming).

A second reason why omitting the random slope tends to be much less consequential in the pure lower-level effect case is the much higher overall precision (expressed for instance in higher absolute *t*-ratios) with which such effects tend to be estimated. Identification of a pure lower-level effect is about estimating the average strength of a lower-level relationship across a set of clusters. Identification of cross-level interactions is about explaining cross-cluster variation in the strength of a relationship. Much more data will usually be needed to gain the statistical power for drawing firm conclusions concerning the second type of effect (Gelman and Hill, 2007: Ch. 20). In consequence, the specification of the random slope term is much less likely to make a difference with respect to conventional levels of significance in the case of pure lower-level effects than in the case of cross-level interactions.

To illustrate this point, we return to the empirical examples from the 'Illustrative Empirical Analyses' section above. In addition to the 60 absolute *t*-ratios for the cross-level interaction estimates (see Figure 2 above and Online Supplement Tables D1–D6), we collected absolute *t*-ratios for the corresponding pure lower-level effects (i.e., the *t*-ratios pertaining to the uninteracted coefficients of high education, intermediate education, gender, unemployment, age, and marital status). The models underlying these *t*-ratios are identical to the ones that underlie Figure 2, with the one exception that they do not include the interactions between the HDI and the lower-level predictors (the additive effect of the HDI remains in the model). As before, we consider two specifications for each of the 30 combinations of lower-level predictors and outcome variables, one that only includes a random intercept and one that additionally includes a random slope term on the lower-level predictor.

Figure 5 shows these absolute *t*-ratios, ranked by their size and differentiated by whether the model entailed a random slope on the respective lower-level predictor. The top graph depicts the *t*-ratios for the cross-level interaction terms, which were already shown in Figure 2 above. The *t*-statistics are mostly smaller than 5 and when a random slope was specified, the vast majority is smaller than the critical value of 2.056 ($df \approx 28 - 2 = 26$, see Elff *et al.*, forthcoming). Because of these generally small *t*-ratios, the inclusion of the random slope term would often lead to qualitatively different conclusions concerning the strength of evidence against the null hypothesis.

The picture looks very different for the absolute *t*-ratios of the pure lower-level effects, displayed in the bottom graph. Including the random slope reduces the distribution of *t*-ratios substantially also in this case. However, the *t*-ratios remain very high and far above conventional thresholds for statistical significance in the vast majority of cases. Of the 26 lower-level effects that are significant at the five per cent level according to a model that omits the respective random slope, 24 remained significant after its inclusion. In the cross-level interaction case, by contrast, we observe a change from statistical significance to insignificance in 7 out of initially 15 cases (see also Figure 2 above). Thus, even though statistical inference for lower-level effects will be over-confident when the corresponding random slope is not included, chances are high that any given effect would remain (highly) significant in the correctly specified model. This is the decisive difference to the cross-level interaction case where switching to the correct specification will often wash away any robust evidence against the null hypothesis.

Conclusions

Our study was motivated by the observation that published research using mixed-effects multilevel models is strikingly inconsistent when it comes to the inclusion of random slopes on the lower-level components of cross-level interactions. Several leading textbooks on multilevel modelling fail to give a clear recommendation on this issue as well.

We have argued, and demonstrated with Monte Carlo simulations, that cross-level interactions generally require the inclusion of the associated random slope. Omission of the random slope term results in unmodelled cluster-driven heteroskedasticity and cluster-correlated errors, thus violating fundamental model assumptions and assuming too much independence among observations. The most important consequence is that statistical inference for the cross-level interaction term and the main effect of its lower-level component becomes overly optimistic: *t*-ratios will be too high, confidence intervals too narrow, and standard errors as well as *P*-values too low, leading to overrejection of the null hypothesis of no effect. The problem becomes more severe (a) as unmodelled variation in the cluster-specific slopes increases, (b) as the variance of the lower-level variable involved in the interaction increases, and (c) as the the cluster size grows (i.e., the more lower-level observations there are per cluster). Mixed-effects models that include a random slope term on the lower-level component of cross-level interaction terms generally performed very well in our simulations. Only in a few situations of little practical relevenace did we find them to produce over-conservative inference.

A total of 30 illustrative applications based on ESS data indicate that the consequences of omitting the random slope can be dramatic in real-life settings. In three quarters of cases, the absolute *t*-statistic on the cross-level interaction term was *at least* 31 per cent lower for the model including the random slope than for the model omitting it. These results are highly discomforting, as our review of *ESR* articles indicates that many published cross-level interactions estimated without the associated random slope are barely statistically significant. It is quite likely that most of these estimates could not be considered as robust evidence for the relationship in question if they were estimated using the correct specification.

The prototypical case that has been guiding our study is that of a cross-sectional cross-country comparison, as this is by far the predominant type of multilevel analysis published in the *ESR*. It is clear, however, that our findings have similar implications for other data structures. An obvious case is that of *repeated* cross-national surveys. Schmidt-Catran and Fairbrother (2016) show that correct inference will typically require the specification of random intercept terms at both the country and the country-year (and, to a lesser extent, also the year) level in this setting. Our findings indicate that it will also be crucial to specify random slope terms at those levels, particularly for lower-level predictors that are involved in cross-level interactions with (contextual) variables that vary at the country or country-year level.[12]

Going beyond the case of cross-national surveys, researchers using mixed-effects models to analyse other types of multilevel data should similarly make sure that their conclusions about cross-level interactions and lower-level effects do not hinge on the omission of the corresponding random slope terms. The consequences may be somewhat less severe when working with meso-level contextual units such as schools or with individual-level panel data (because average cluster size tends to be lower). Nevertheless, failing to specify a random slope has the potential to compromise statistical inference also in these settings.

Looking backward, our results thus cast doubt on many findings that are potentially considered well-established. We encourage researchers to take our results into account when reviewing previous studies. Results on cross-level interactions that were estimated without the crucial random slope term should be interpreted with caution and considered as preliminary. Their validity should be checked through replication, and the results of replication attempts should be publicly reported to promote a balanced assessment of the empirical evidence for a given cross-level relationship.

Looking forward, our findings suggest that researchers who investigate cross-level interactions using mixed-effects multilevel models should *always* include a random slope for the lower-level component of the interaction. Editors and referees should insist that authors adhere to this rule.

Last but not least, our results highlight another, broader challenge faced by those who want to analyse multilevel data with mixed-effects models. We found that random slopes are similarly required for accurate inference about ‘pure’ lower-level effects, provided—of course—that the effect truly varies across clusters (see also Barr *et al.*, 2013; Bell, Fairbrother and Jones, 2018). We believe this issue to be less troubling than the cross-level interaction case because researchers using multilevel modelling are rarely interested in pure lower-level effects and because many of these effects would remain highly statistically significant even if the associated absolute *t*-statistic declined by 50 per cent or more. Nevertheless, the idea that statistical inference on lower-level predictors will typically be anti-conservative is unattractive, even if they are usually only considered as control variables.

How, then, can this issue be resolved? Simply specifying random slopes on all lower-level predictors will rarely be a solution. Such models would typically suffer from overspecification (Bates *et al.*, 2015; Heisig, Schaeffer and Giesecke, 2017; Matuschek *et al.*, 2017). The strategy of specifying additional random slopes in the interest of accurate inference would quickly become self-defeating, leading to the very problem it seeks to solve: anti-conservative inference (Heisig, Schaeffer and Giesecke, 2017). One viable, albeit not fully satisfactory, solution will be to focus on achieving correct inference for the coefficients of interest and take inference for other predictors with a large grain of salt. One might also consider fitting the same fixed-effects specification (i.e., the same model in terms of the set of predictors) with several random-effects specifications, including the random slope terms one at a time (i.e., first for x_1, then for x_2, and so forth) to get a sense of the correct standard errors for the different lower-level predictors. A fully convincing solution will probably require approaches such as bootstrapping or profile likelihood methods, however. Methods for cluster-robust inference from the econometric literature may also be worth considering, but they face their own set of challenges.[13]

Notes

[1] The careful reader might notice that Equations 5 and 6 in Snijders and Bosker (2012) refer to the

conditional variance of the outcome Y_{ij} rather than the overall error v_{ij}. However, this is fully consistent with the formulation given here because conditional on x_{ij} variation in Y_{ij} can only come from the random part of the model, that is, from v_{ij}.

2 We do not study the performance of cluster-robust methods in this article because mixed-effects models are by far the most widely used method for investigating context effects in sociology (Heisig, Schaeffer and Giesecke, 2017) and because the cluster-robust approach has its own set of pitfalls, especially when the number of clusters is small or when the data are characterized by multiple (non-hierarchical) levels of clustering (for further discussion, see Cameron and Miller, 2015).

3 This is not to say that point estimates will never differ according to whether a random slope is included or not. This is easiest to see in the case of a 'pure' lower-level effect, that is, of a coefficient of a lower-level variable that is not interacted with an upper-level predictor. In the model that includes the random slope, the coefficient estimate on the lower-level predictor is an estimate of the *unweighted* average of the cluster-specific slopes. This follows from the fact that the random slope is assumed to be normally distributed with a mean of 0. In the model that does not include a random slope, the coefficient will be a *weighted* average of the cluster-specific coefficients. Therefore, the difference will be particularly large when the magnitude of the cluster-specific coefficients is strongly related to cluster size. It is not clear whether one would necessarily want to describe one of these estimates as 'biased', however, as the two approaches really estimate different quantities. To see that similar issues arise in the estimation of cross-level interactions, one simply has to note that the coefficient on the cross-level interaction term can be conceptualized as the effect of the cluster-level variable on the *conditional* average slope of the lower-level variable. Equation 3 makes this very clear.

4 We have conducted additional Monte Carlo simulation results that support this claim. These results are available upon request.

5 Barr *et al.* (2013) also stress the importance of random slope terms for statistical inference, but they focus on experimental designs with crossed random effects that are quite different from those typically encountered in sociology.

6 The simulation results indeed show that both types of models produce unbiased coefficient estimates. These results can be obtained from the replication files that are part of the online supporting material. As discussed in footnote 3, there may be cases when a model with and a model without a random slope produce systematically different estimates, but the reason here would be that the former estimates an unweighted whereas the latter estimates a weighted average effect.

7 In other words, while the nominal probability of committing a Type 1 error, that is, of rejecting the null hypothesis of no effect although it is true, would be 0.05, the true probability would be 0.10.

8 Replication code for the analyses in Heisig, Schaeffer and Giesecke (2017) is available at http://journals.sagepub.com/doi/suppl/10.1177/0003122417717901. Together with the replication code for the present article, it can be used to replicate all analyses reported in this section.

9 When relying on the *P*-value, we assumed a normally distributed test statistics, consistent with the approach taken by the majority of authors. Elff *et al.* (forthcoming) show this assumption to be problematic when the number of clusters is small, but we nevertheless use it here to treat the different studies consistently.

10 For a thorough review and critical discussion of reporting practices and significance testing in the *ESR*, see Bernardi, Chakhaia and Leopold (2017).

11 We focus on statistical significance because of the important role that it continues to play in the publication process and in the evaluation of empirical evidence. We do not mean to imply that statistical significance is the best and/or should be the only criterion used to assess statistical uncertainty. Our conclusions would clearly be similar for alternative measures of uncertainty such as standard errors or confidence intervals.

12 One might wonder why models without a random slope term performed well in Schmidt-Catran and Fairbrother's (2016) Monte Carlo simulation. The reason is that the DGP underlying their simulations did not involve any random slopes, so that their omission did not result in inferential deficiencies.

13 Conventional corrections—as implemented in Stata's vce (cluster *clustvar*) option—are known to require a substantial number of clusters (at least 40 or 50) for accurate inference (Cameron and Miller, 2015). Recent methods for the few-clusters case show promising performance, but further research is needed before clear recommendations can be given (for details, see Cameron and Miller, 2015; Esarey and Menger, 2018). In addition, cluster-

robust methods treat the violation of classical assumptions as a mere nuisance. There may be substantial benefits to addressing these violations through model respecification, for example, through the inclusion of additional predictors or random slope terms that capture heterogeneous effects (King and Roberts, 2015; Heisig, Schaeffer and Giesecke, 2017).

Supplementary Data

Supplementary data are available at *ESR* online.

Acknowledgements

Parts of this article were presented at the RC 28 Spring Meeting 2018, the DGS-Kongress 2018, and the ECSR 2018 Conference. The authors thank participants for their feedback. The authors are particularly indebted to two anonymous *ESR* reviewers for helpful comments and to Mark Wittek for thoroughly coding hundreds of cross-level interactions.

References

Barr, D. J. *et al*. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language*, **68**, 255–278.

Bates, D. *et al*. (2015). Parsimonious mixed models. *arXiv*, preprint arXiv:1506.04967. https://arxiv.org/abs/1506.04967

Bates, D. *et al*. (2017). lme4: linear mixed-effects models using Eigen and S4. R Package Version 1.1-15, https://CRAN.R-project.org/package=lme4.

Bell, A., Fairbrother, M. and Jones, K. (2018). Fixed and random effects models: making an informed choice. *Quality and Quantity*. https://doi.org/10.1007/s11135-018-0802-x

Berkhof, J. and Kampen, J. K. (2004). Asymptotic effect of misspecification in the random part of the multilevel model. *Journal of Educational and Behavioral Statistics*, **29**, 201–218.

Bernardi, F., Chakhaia, L. and Leopold, L. (2017). 'Sing me a song with social significance': the (mis) use of statistical significance testing in European sociological research. *European Sociological Review*, **33**, 1–15.

Bryan, M. L. and Jenkins, S. P. (2016). Multilevel modelling of country effects: a cautionary tale. *European Sociological Review*, **32**, 3–22.

Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, **50**, 317–372.

Elff, M. *et al*. (forthcoming). Multilevel analysis with few clusters: improving likelihood-based methods to provide unbiased estimates and accurate inference. *British Journal of Political Science*.

Esarey, J. and Menger, A. (2018). Practical and effective approaches to dealing with clustered data. *Political Science Research and Methods*, 1–19. doi:10.1017/psrm.2017.42.

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

Franco, A., Malhotra, N. and Simonovits, G. (2014). Publication bias in the social sciences: unlocking the file drawer. *Science*, **345**, 1502–1505.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*, Cambridge: Cambridge University Press.

Heisig, J. P., Schaeffer, M. and Giesecke, J. (2017). The costs of simplicity: why multilevel models may benefit from accounting for cross-cluster differences in the effects of controls. *American Sociological Review*, **82**, 796–827.

King, G. and Roberts, M. E. (2015). How robust standard errors expose methodological problems they do not fix, and what to do about it. *Political Analysis*, **23**, 159–179.

Matuschek, H. *et al*. (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language*, **94**, 305–315.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing.

Rabe-Hesketh, S. and Skrondal, A. (2012). *Multilevel and Longitudinal Modeling Using Stata. Volume I: Continuous Responses*, 3rd edn. College Station, TX: Stata Press.

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage.

Schmidt-Catran, A. W. and Fairbrother, M. (2016). The random effects in multilevel models: getting them wrong and getting them right. *European Sociological Review*, **32**, 23–38.

Simonsohn, U., Nelson, L. D. and Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of Experimental Psychology: General*, **143**, 534–547.

Simonsohn, U., Nelson, L. D. and Simmons, J. P. (2015). Better P-curves: making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a reply to Ulrich and Miller. *Journal of Experimental Psychology: General*, **144**, 1146–1152.

Snijders, T. and Bosker, R. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.

Student (1908). The probable error of a mean. *Biometrika*, **6**, 1–25.

Jan Paul Heisig heads the research group 'Health and Social Inequality' at WZB Berlin Social Science Center. His research focuses on social inequalities in health, education, and labor market attainment as well as quantitative methods. Recent articles have appeared in journals

such as *American Sociological Review*, *Sociology of Education*, and *Ageing & Society*. His book '*Late-career Risks in Changing Welfare States*' was published by Amsterdam University Press in 2015.

Merlin Schaeffer is Associate Professor at the Department of Sociology, University of Copenhagen. His research interests include the comparative analysis of international migration and immigrant integration, social stratification, political sociology, and quantitative methods. His recent work has appeared in *American Journal of Sociology*, *European Sociological Review*, and *American Sociological Review*. His monograph '*Ethnic Diversity and Social Cohesion*' was published by Routledge in 2015.