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

Heisig, Jan Paul; Schaeffer, Merlin

Published in:
European Sociological Review

DOI:
10.1093/esr/jcy053

Publication date:
2019

Document version
Early version, also known as pre-print

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

Jan Paul Heisig
WZB Berlin Social Science Center

Merlin Schaeffer
University of Copenhagen

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. Monte Carlo simulations and illustrative empirical analyses highlight the practical relevance of the issue. Using European Social Survey data, we examine a total 30 cross-level interactions. Introducing a random slope term on the lower-level variable involved in a cross-level interaction, reduces the absolute t-ratio by 31% or more in three quarters of cases, with an average reduction of 42%. 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 meet conventional standards of statistical significance if estimated using the correct specification. This raises the question how much robust evidence of cross-level interactions sociology has actually produced over the past decades.

*Parts of this paper were presented at the RC 28 Spring Meeting 2018. We thank participants for their feedback. We are particularly indebted to Mark Wittek for thoroughly coding hundreds of cross-level interactions. Both authors have contributed equally. Please direct correspondence to Merlin Schaeffer, University of Copenhagen, Department of Sociology, Øster Farimagsgade 5, Building 16, DK-1353 København K.

One of the enduring questions of sociology is how human attitudes and behavior 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 neighborhood 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 modeling 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 to 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 et al., 2017) and for statistical inference (Barr et al., 2013; Bell et al., 2016; Berkhof and Kampen, 2004; Bryan and Jenkins, 2016; Schmidt-Catran and Fairbrother, 2016).

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 put everyone in danger of crime. Or to put it another way, the better-educated tend to benefit the most from improving societal conditions, whereas the less educated remain relatively vulnerable to crime. 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%) 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 modeling 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 autocorrelation related to the lower-level component of the cross-level interaction (and hence to the cross-level interaction term itself). As in standard linear regression, ignoring heteroskedasticity and autocorrelation 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 a 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 in-

ference for the cross-level interaction term and the main effect of its lower-level component. The proportion of 95% 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 (unmodeled) 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 t-ratio on the cross-level interaction term by 31% or more in three quarters of cases, with an average reduction of 42%.

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 et al., 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 whereas the t-ratios for models that include a random slope do not. In combination with the results of our Monte Carlo simulations and empirical illustrations, our review therefore suggests that most 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 coun-

tries 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, 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% 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 (country-)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 that they 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, Rabe-Hesketh and Skrondal, 2012; Raudenbush and Bryk, 2002; Snijders and Bosker, 2012). We begin with the following lower-level equation for the (lower-level) outcome Y_{ij} (e.g., fear of crime):

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

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 unmodeled 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 to 3 highlight the multilevel nature of the model. An alternative formulation is the 'expanded form' of the model, derived by substituting Equations 2 and 3 into 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 for 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 section 'Illustrative Empirical Analyses' and in the Online Supplement (see, in particular, the final columns of Table D1 to 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 variation in $\beta_j^{(x)}$. We find that inference can be substantially overoptimistic even when $R^2(\beta_j^{(x)})$ of the implicit cluster regression is as high as .95 or when standard model selection criteria such as likelihood ratio tests or information criteria indicate that the remaining variation is negligible and favor the model that drops the random slope (the results on model selection strategies can be found in Appendix C in the Online Supplement).

The two-stage formulation of the model in Equations 1 to 3 also suggests that omission of $u_j^{(x)}$ should primarily affect inference about $\gamma^{(x)}$ and $\gamma^{(xz)}z_j$ 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 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]

$$\text{Var}(v_{ij}|x_{ij}) = \text{Var}(u_j^{(c)}) + 2\text{Cov}(u_j^{(c)}, u_j^{(x)})x_{ij} + \text{Var}(u_j^{(x)})x_{ij}^2 + \text{Var}(\epsilon_{ij}). \quad (5)$$

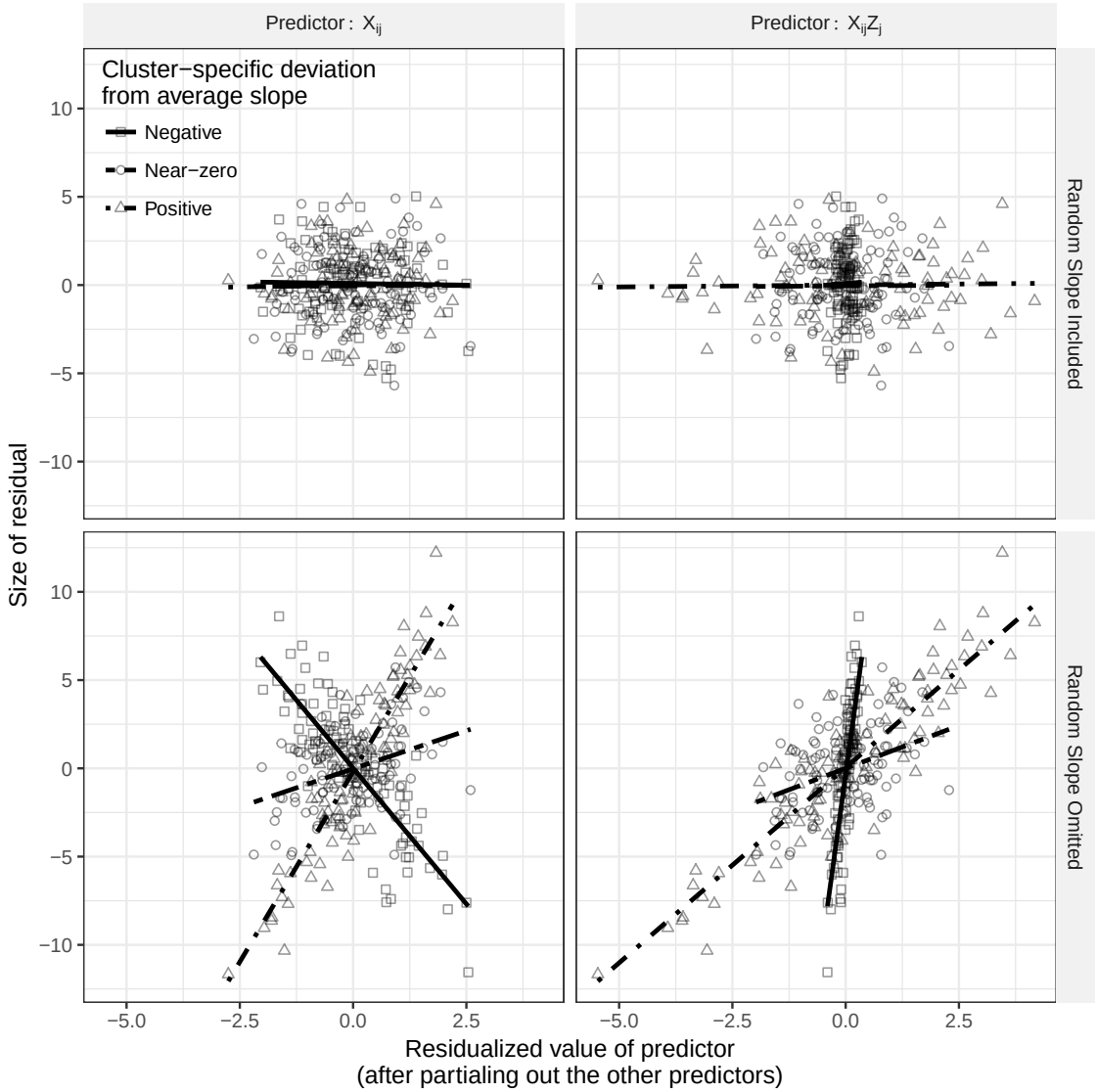
The covariance of the error terms for two different individuals (say, i and i') belonging to the same cluster will be (Snijders and Bosker, 2012, Equation 5.6):

$$\text{Cov}(v_{ij}, v_{i'j}|x_{ij}, x_{i'j}) = \text{Var}(u_j^{(c)}) + \text{Cov}(u_j^{(c)}, u_j^{(x)})(x_{ij} + x_{i'j}) + \text{Var}(u_j^{(x)})x_{ij}x_{i'j}. \quad (6)$$

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) unmodeled heteroskedasticity in the error term (due to the second and third term on the right hand side in Equation 5) and b) unmodeled 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 to 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

Figure 1: Lower-level Residuals for Models with and without Random Slope



Note: Residuals are from linear mixed effects models. The data are are simulated according to Equations 1 to 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 .6 and 2, respectively, and with a correlation of .3; the lower-level error ϵ_{ij} is normally distributed with a standard deviation of 2.

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 partialing out the cluster-level predictor z_j. We focus on three representative clusters, one with a slope for $\beta_j^{(x)}$ that deviates strongly positive from the average slope, one with a slope for $\beta_j^{(x)}$ that is close to the average, and one with a slope for $\beta_j^{(x)}$ that deviates strongly negative 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 left-column 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 right column. 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 (auto-)correlated when they have similar values on x_{ij} and $z_j x_{ij}$.

Omitting a random slope that actually belongs in the model thus leads to unmodeled heteroskedasticity and autocorrelation. 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 modeling 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 autocorrelation is a common motivation for both multilevel modeling 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 to 6 above). The latter uses special 'sandwich-

type' estimators of the coefficient covariance matrix that remain consistent even in the presence of heteroskedasticity and (within-cluster) autocorrelation.

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 $Var(u_j^{(x)})$. The second factor is the extent of variation in the lower-level predictor, that is, Var$(x_j^{(x)})$. As Var$(x_j^{(x)})$ increases, so will the extent of (unmodeled) 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, 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.

Another way to understand why the average cluster size matters is to consider how it affects the overall prevalence of autocorrelation in the data (see Cameron and Miller, 2015, 319ff.). Both the multilevel and the cluster-robust approach assume that errors are correlated within, but not across clusters. Thus off-diagonal elements of the error covariance matrix will be zero for pairs of observations that belong to different clusters and will generally be non-zero for pairs of observations that belong to the same cluster. For a given lower-level sample size, this means that the number of non-zero off-diagonal elements increases with the average cluster size (and decreases with the number of clusters).[3]

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 autocorrelation, 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.[4] 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.[5] 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 (cf., Barr et al., 2013; Bell et al., 2016). 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? After all, it is the presence of an unmodeled 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 autocorrelation (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.

That 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 crucial reason for this is that, compared to a 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 evi-

dence against the null hypothesis in the cross-level interaction case (say, between $p < .05$ and $p > .1$). In the case of pure lower-level effects, the difference is more likely to be between different degrees of strong evidence (say, between $p < .001$ and $p < .01$). We explore these issues in detail in the 'Random Slopes and 'Pure' Lower-level Effects' section. In a first step, we now focus on 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 data-generating process (DGP) and applying the estimators to each simulated dataset. 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 analyze

Table 1: Dimensions Manipulated in the Monte Carlo experiments

Dimension	Levels
m	5
Number of clusters	15
	25
n_g	100
Number of observations	500
per cluster	1000
SD($u_j^{(x)}$) and $R^2(\beta_j^{(x)})$	.1005 (.99)
Standard deviation of	.1429 (.98)
random slope term $u_j^{(x)}$	.2294 (.95)
(implied cluster-level $R^2(\beta_j^{(x)})$	.3333 (.90)
for cluster-level	1.0000 (.50)
regression in parentheses)	3.0000 (.10)
SD(x_{ij})	.50
Standard deviation of	1.00
lower-level predictor x_{ij}	2.00

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.

We obtain 10,000 replications (i.e., 10,000 simulated datasets) 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. The second model omits this term. 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 the random slope on the lower-level component of the cross-level interaction (see also the 'Cross-level Interactions in the ESR' section below).

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% con-

fidence intervals. Accurate inference (for an unbiased estimator) requires that the actual coverage rate equals the nominal rate. We therefore examine whether two-sided 95% confidence intervals cover the true parameter in more or less than 95% of the 10,000 Monte Carlo replications. Let $C_{95}(r) = 1$ if the two-sided 95% confidence interval for the r^{th} replication includes the true value of the parameter of interest and zero otherwise. Then coverage is defined as

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

If coverage is greater than 95%, 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%, 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 to 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 (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., 2016, to select the appropriate t-distribution). We further explore these issues and present results on standard error optimisim in Online Supplement Appendix 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 ($= \sqrt{.95 \times .05/10000}$) or .218 percentage points. Thus, the actual coverage rate of an estimator (for a given experimental condition) is significantly different (at the five percent level) from the nominal level of 95% if it deviates from that level by more than .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. (2016), 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 not apply to inference for the main effect of the contextual predictor z_j where coverage rates fall within the range of $95 \pm 0.427\%$ 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

Table 2: Actual Coverage Rates of Nominal 95% Confidence Interval by Variance of Lower-level Predictor and Random Slope Term

	$\gamma^{(x)}$		$\gamma^{(xz)}$		$\gamma^{(cz)}$	
	Random Slope		Random Slope		Random Slope	
SD(x_{ij})	Included	Omitted	Included	Omitted	Included	Omitted
$R^2(\beta_j^{(x)}) = 0.95$ (i.e., SD($u_j^{(x)}$) ≈ 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($u_j^{(x)}$) ≈ 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

Note: Results are based on 10,000 Monte Carlo replications. Because of Monte Carlo sampling error, the 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.

an actual coverage rate of 90% means that nominal significance on the 5% level would actually only mean 'marginal' significance on the 10% level.[7] Yet, most actual coverage rates displayed in Table 2 are even substantially smaller than 90%. 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.

By contrast, coverage rates of confidence intervals based on models that include a 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. 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})). The reason is that the extent of heteroskedasticity and autocorrelation that remains unmodeled in the specification that omits the random slope is a function of the product of these two factors (see Equations 5 and 6 above), which 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 $R^2(\beta_j^{(x)})$

Table 3: Actual Coverage Rates (%) of Nominal 95% Confidence Interval by Number of Clusters and Lower-level Observations

		$\gamma^{(x)}$		$\gamma^{(xz)}$		$\gamma^{(cz)}$	
		Random Slope		Random Slope		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	2500	95.09	43.23	95.07	43.68	93.64	95.34
1000	5000	95.07	31.39	94.58	31.70	93.95	95.11
				$m = 15$ Clusters			
100	1500	95.19	58.57	94.75	58.62	93.65	95.51
500	7500	94.90	30.55	94.84	30.51	94.80	95.19
1000	15000	94.93	21.46	94.95	22.37	95.10	95.25
				$m = 25$ Clusters			
100	2500	94.79	56.87	95.24	56.22	93.23	95.03
500	12500	94.93	29.71	94.98	29.29	95.13	95.14
1000	25000	94.85	21.43	95.23	21.32	94.90	94.74

Note: Results are based on 10,000 Monte Carlo replications. Because of Monte Carlo sampling error, the test interval is 95 ± 0.427. Values smaller or larger than that are statistically significant deviations and indicate biased inference. These results are based on experimental conditions for which $R^2(\beta_j^{(x)}) = 0.50$ (i.e., $\text{SD}(u_j^{(x)}) = 1$), and $\text{SD}(x_{ij}) = 1$.

to .5 (i.e., $\text{SD}(u_j^{(x)}) = 1.00$); 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. As befoe, we also see that omitting the random slope term does not, in general, compromises inference for $\gamma^{(cz)}$. As expected, the problem gets worse as the cluster size (i.e., the number of lower-level observations per cluster) increases. For every given number of clusters, undercoverage is most severe for 1000 observations per cluster as compared to 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, under-coverage increases with the extent of variation in the lower-level variable, the

extent of variation in the unmodeled random slope term, and with the (average) size of the clusters.

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

In Online Supplement Appendix B we further investigate the unexpected result that the (correctly specified) model including the random slope term yields overconservative statistical inference in some situations. We present additional simulations that consider even lower values of .14 and .10 for the standard deviation of the random slope term, implying values of .98 and .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 Table D1 to D6). Moreover, practitioners can easily verify if they are dealing with a situation where the random slope variation is close to zero.

In a second set of supplementary analyses, presented in Online Supplement Appendix 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 analyses based on European Social Survey data (ESS Round 6, 2016).

We adopt Heisig et al.'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 education, age, gender, unemployment, being married, and having a medium education) and five standard outcome variables (generalized trust, homophobia, xenophobia, fear of crime, and occupational status) are moderated by the Human Development Index (HDI).

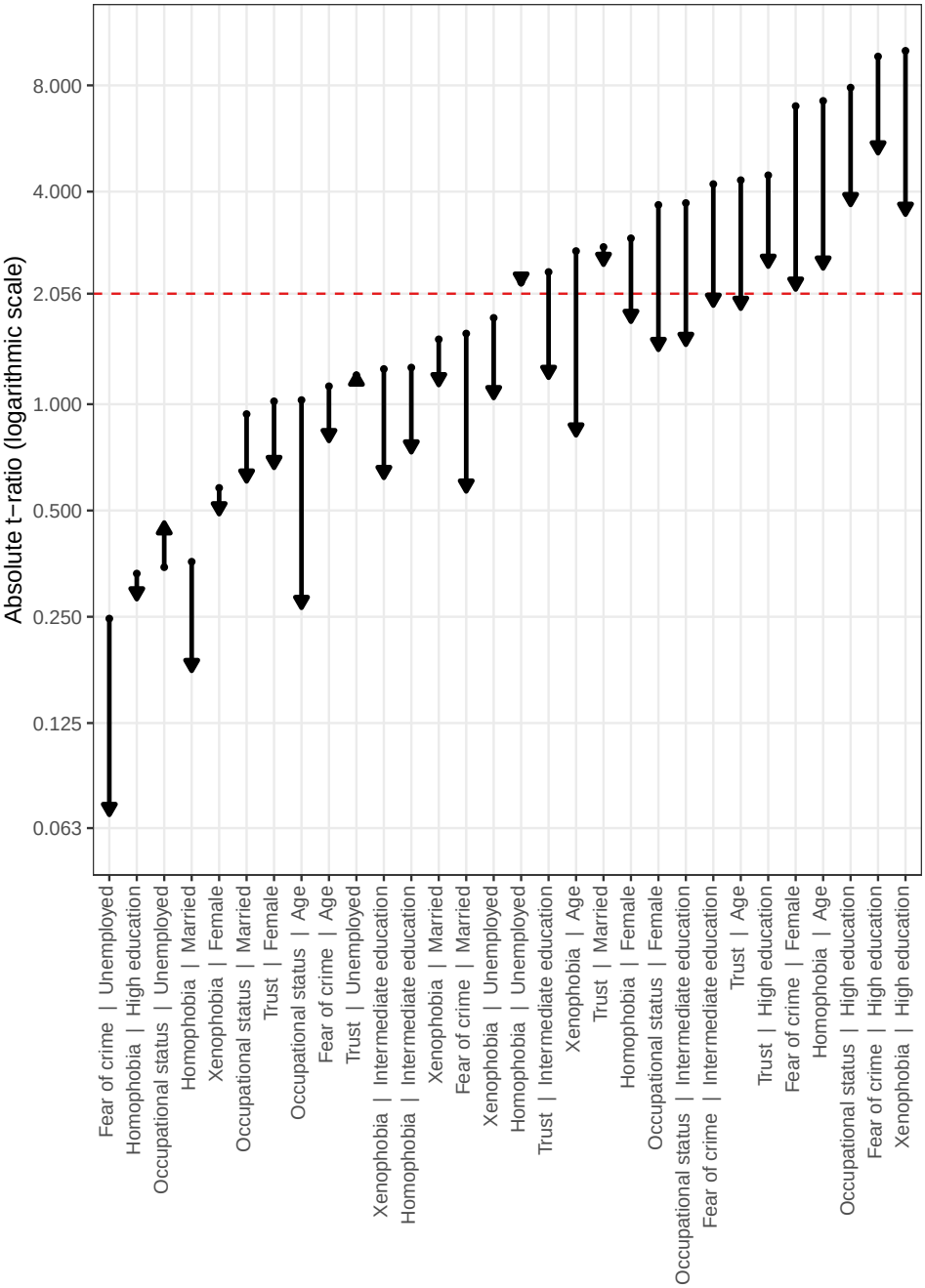
For each of the 30 cross-level interactions (5 dependent variables $\times$ 6 lower-level predictors) we estimate two specifications, resulting in a total of 60 linear mixed effects models. The first specification is a random intercept and slope model that assigns a random effect to the coefficient of the lower-level variable involved in the particular 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. Appendix D in the Online Supplement gives a brief description of the coding of the variables and provides exact results for the coefficients of interest in Tables D1 to D6. For brevity, we focus on statistical inference for the cross-level interaction term in the main article. In line with our Monte Carlo simulation results, Tables D1 to D6 suggest similar conclusions for the main effect of the lower-level predictor and no consequences of omitting the random slope term for 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 omitting the random slope is only 5.1, a reduction of 46.8% (see Online Supplement D1 for these values and the associated point estimates). Figure 2 shows that reductions of such alarming magnitude are not the exception. Note that the y-scale is logged, so 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%

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



Note: The triangled arrow heads shows 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 indicate 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 human development index. The overall 60 for cross-level interactions are estimated by linear mixed effects models which are displayed in Tables D1 to D5. The dashed horizontal line demarcates 2.056, the threshold for statistical significance at the five percent 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. (2016).

on average. The median reduction is 48.3% and the 25th and 75th percentiles are 31.3 and 60.9%, respectively.

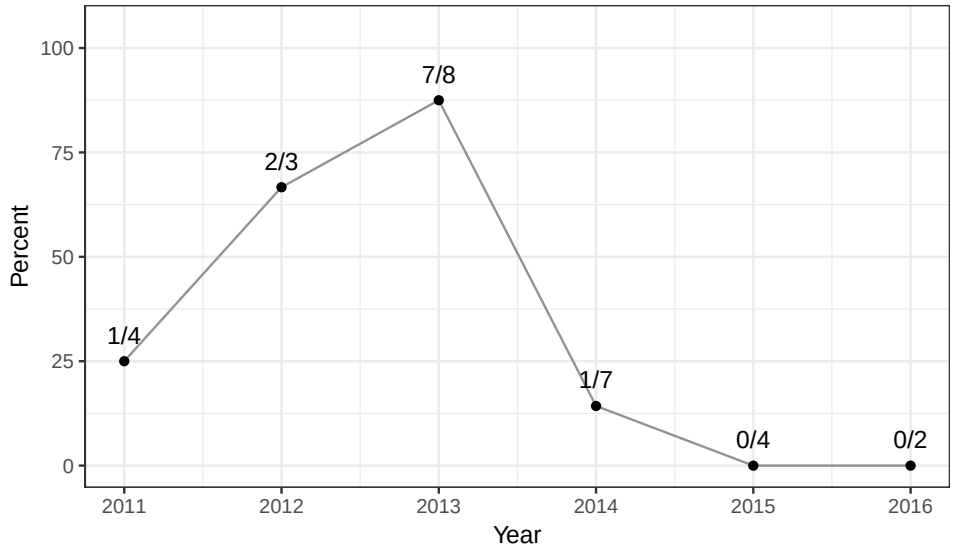
The final columns of Online Supplement Tables D1 to D6 convey another important result. They display the remaining variation of the random slope in the model including the cross-level interaction, expressed as the ratio of the random slope standard deviation to the corresponding main effect. Thus, the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

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 modeling 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. For simplicity, we confined ourselves to studies using simple two-level models where lower-level observations are nested in exactly one type of upper-level unit. Overall we identified 28 studies, the vast majority of which (24 or 86%) were country comparisons (one of the remaining studies treated individuals as nested in combinations of countries

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.

and survey years). The 28 studies reported a total of 150 estimates of cross-level interactions. When a paper provided multiple estimates of the same cross-level interaction (e.g., for different model specifications or subsamples), we chose one estimate at random. Note that we now disregard the main effects of the cluster- and lower-level components 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%) specify random slopes on the lower-level components of the cross-level interactions they investigate. Figure 3 displays the percentage of studies that include random slope terms by year of publication. It provides no evidence that correct specifications have become more popular over time; if anything, the contrary seems true. Since 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.

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

		Random Slope	
		Included	Omitted
Insignificant	($p >= 0.1$)	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 ($n = 150$)		($n = 51$)	($n = 99$)

Note: Results are based on 28 articles reporting 150 cross-level interactions from two-level mixed effects models published in the ESR 2011-2016. Since 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.

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% 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.

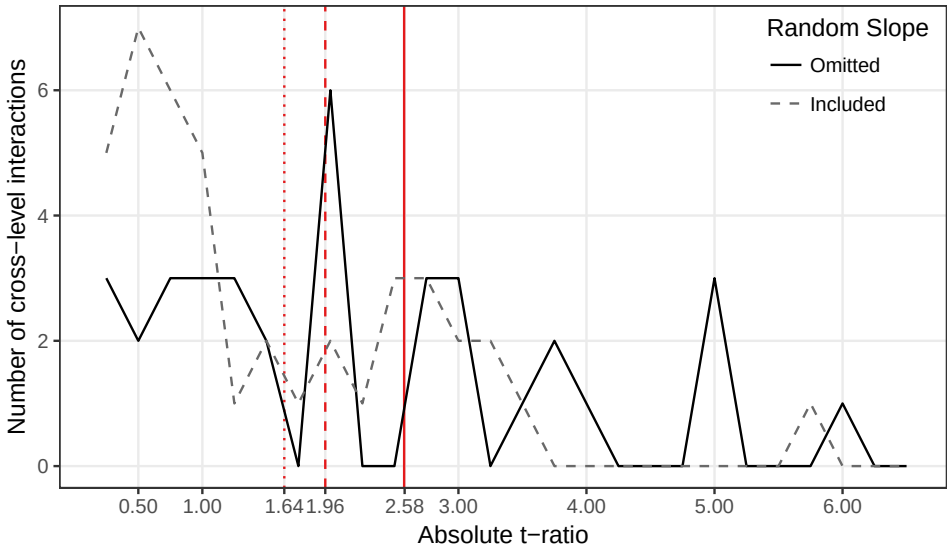
Table 4 displays the percentage of estimated cross-level interaction effects that attain a given level of statistical significance according to whether the model did or did not include a random slope on the lower-level component. It shows a consistent pattern of more insignificant and marginally significant, but fewer significant and highly significant cross-level interaction terms for models with the

correct random effects specification (i.e., models that include a random slope). Put differently, cross-level interactions that are erroneously estimated without a random slope on the lower-level component more often reach (higher levels of) statistical significance. 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. Proving this is difficult, since about 60% of null-results are never written up (Franco et al., 2014). Online Supplement Appendix E uses p-curve analyses following Simonsohn et al. (2014, 2015) to investigate the possibility of publication bias and 'p-hacking' more systematically.

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 outputs—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 black solid line shows the distribution of t-ratios from misspecified models that omit the crucial random slope term. The gray dashed line shows the distribution from models that include it.

Figure 4 shows a pronounced peak near the threshold for statistical significance at the 5% ($t = 1.96$) level. This unnatural peak characterize the distribution

Figure 4: Distribution of absolute t-ratios of cross-level interactions



Note: Results are based on 86 cross-level interaction terms from two-level mixed effects models for which the authors reported exact inference statistics. These were reported in 20 articles published in the ESR 2011-2016. Bin width is set to 0.25.

of t-ratios especially for the incorrectly specified models and is suggestive of p-hacking. Online Supplement Appendix 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 solid line for cross-level interactions tested without a random slope needs to shift 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 analyses at face value, the correct t-ratios will be at least 31% 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., $t = 1.96$ and $t = 2.58$) are questionable when cluster-level samples are small. Elff et al. (2016) 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 autocorrelation (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 coefficient 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 $\gamma_j^{(x)}$. 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 overconfidence, as measured by the undercoverage of two-sided 95% confidence intervals, is generally very similar to the corresponding results for the cross-level interaction case.

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 analyze 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 composi-

Table 5: Actual Coverage Rates of Nominal 95% Confidence Interval by Variance of Lower-level Predictor and Random Slope Term

	$\gamma^{(x)}$	
	Random Slope	
$SD(x_{ij})$	Included	Omitted
$SD(u_j^{(x)}) \approx 0.23$		
0.5	96.43	93.16
1.0	95.60	81.58
2.0	95.17	56.26
$SD(u_j^{(x)}) \approx 0.33$		
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

Note: Results are based on 10,000 Monte Carlo replications. Because of Monte Carlo sampling error, the 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.

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

Note: 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.

tional differences among clusters. So while inference for lower-level effects might be overconfident, 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 a) the title, b) the abstract, and (if existent) c) explicitly formulated hypotheses stress 1) individual-level relationships, 2) contextual relationships (direct context effects and/or cross-level interactions), or 3) 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%) using (two-level) mixed effects models exclusively posit hypotheses about lower-level effects. By contrast, 53.6% formulate hypothesis about both pure lower-level and contextual relationships and 41.1% only present hypotheses about contextual 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

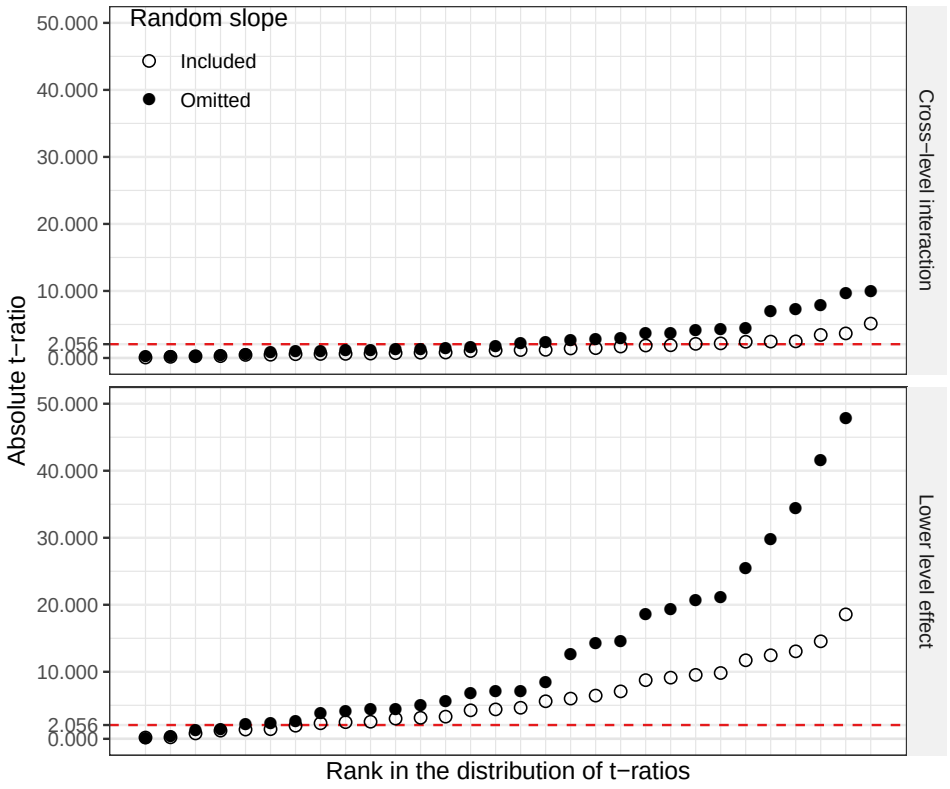
are forced to stress the cardinal contribution of their paper, the mixed category shrinks to ca. 15%—mostly because articles tend to highlight only context effects in their title. Two thirds of all articles fall into this category.

A second reason why omitting the random slope tends to be much less consequential for 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, considering random slopes or not will rarely make a difference with respect to conventional levels of significance in the case of pure lower-level effects.

To illustrate this point, we collected the absolute t-ratios for the 60 cross-level interaction estimates reported in Online Supplement Tables D1 to D6 (5 dependent variables $\times$ 6 lower-level predictors $\times$ 2 specifications, the one including and the one omitting the random slope). In addition, we estimated the same models without the cross-level interaction terms and collected the absolute t-ratios of the 60 pure lower-level effects (i.e., the absolute t-ratios pertaining to the uninteracted coefficients of high education, intermediate education, gender, unemployment, age, and marital status).

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 or not. The t-ratios for the cross-level interaction terms, displayed in the top graph, are mostly smaller than 5 and if a random slope was specified, the vast majority is smaller than the critical value 2.056 (df ≈ 26, see Elff et al. 2016). Because of these generally small t-ratios, the inclusion of the random slope term

Figure 5: Distribution of absolute t-ratios



Note: The 60 absolute t-ratios for cross-level interactions are estimated by linear mixed effects models which are displayed in Tables D1 to D5. The 60 absolute t-ratios for lower-level effects are estimated by identical linear mixed effects models which 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 percent 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. (2016).

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. 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 percent 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 overconfident 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 modeling 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 unmodeled cluster-driven heteroskedasticity and autocorrelation, thus violating fundamental model assump-

tions 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 unmodeled 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 did we find them to produce over-conservative inference, but these issues arose only under conditions of little practical relevance.

A total of 30 illustrative applications based on European Social Survey 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% lower for the model including the random slope than for the model omitting it. These results are highly discomforting since 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.

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.

That being said, our results highlight another, broader challenge faced by those who want to analyze 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 et al., 2016). We believe this issue to be less troubling than the cross-level interaction case because researchers using multilevel modeling 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% 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, an issue discussed in great detail by Bates et al. (2015) and Heisig et al. (2017). The strategy of specifying additional random slopes in the interest of accurate inference would at some point become self-defeating, leading to the very problem it seeks to solve: anti-conservative inference (Heisig et al., 2017). One viable, albeit not fully satisfactory, solution will be to focus on achieveing 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 predic-

tors included) 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, however, will probably require methods such as bootstrapping, profile likelihood methods, or generalizations of 'sandwich-type' methods for cluster-robust inference (as implemented, for example, in the `vce(cluster clustvar)` option for Stata's `mixed` command).[12] There is good justification for all of the latter methods, but further research is needed to learn about their peformance in typical sociological applications that often involve a limited number of large clusters and many lower-level controls.

References

Barr, D. J., Levy, R., Scheepers, C. and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, **68**, 255–278.

Bates, D., Kliegl, R., Vasishth, S. and Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bates, D., Maechler, M., Bolker, B. and Walker, S. (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. (2016). Fixed and Random effects models: making an informed choice. *Working Paper*.

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, ISSN 1076-9986.

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, ISSN 0266-7215.

Bruns, S. B. and Ioannidis, J. P. A. (2016). p-Curve and p-Hacking in Observational Research. *PLOS ONE*, **11**, e0149144, ISSN 1932-6203.

Bryan, M. L. and Jenkins, S. P. (2016). Multilevel Modelling of Country Effects: A Cautionary Tale. *European Sociological Review*, **32**, 3–22, ISSN 0266-7215, 1468-2672.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, **33**, 261–304, ISSN 0049-1241.

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., Heisig, J. P., Schaeffer, M. and Shikano, S. (2016). No Need to Turn Bayesian in Multilevel Analysis with Few Clusters: How Frequentist Methods Provide Unbiased Estimates and Accurate Inference. *SocArXiv/Open Science Framework*.

Esarey, J. and Menger, A. (2018). Practical and Effective Approaches to Dealing With Clustered Data. *Political Science Research and Methods*, first View (published online, Jan 19, 2018).

ESS Round 6, E. S. S. (2016). *ESS-6 2012 Documentation Report. Edition 2.2*, 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, ISSN 0036-8075, 1095-9203.

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

Giesbrecht, F. G. and Burns, J. C. (1985). Two-Stage Analysis Based on a Mixed Model: Large-Sample Asymptotic Theory and Small-Sample Simulation Results. *Biometrics*, **41**, 477, ISSN 0006341X.

Grotenhuis, M. t., Pelzer, B., Eisinga, R., Nieuwenhuis, R., Schmidt-Catran, A. and Konig, R. (2016). When size matters: advantages of weighted effect coding in observational studies. *International Journal of Public Health*, 1–5, ISSN 1661-8556, 1661-8564.

Gurland, J. and Tripathi, R. C. (1971). A Simple Approximation for Unbiased Estimation of the Standard Deviation. *The American Statistician*, **25**, 30–32.

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.

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*, third edition edn., College Station, Tex.: Stata Press.

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

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, **2**, 110–114, ISSN 00994987.

Schaalje, G. B., McBride, J. B. and Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics*, **7**, 512–524, ISSN 1085-7117, 1537-2693.

Schmidt-Catran, A. W. and Fairbrother, M. (2015). The Random Effects in Multilevel Models: Getting Them Wrong and Getting Them Right. *European Sociological Review*, **Doi: 10.1093/esr/jcv090**.

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, ISSN 0266-7215, 1468-2672.

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, ISSN 1939-2222(Electronic),0096-3445(Print).

Simonsohn, U., Simmons, J. P. and Nelson, L. D. (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. A. B. and Bosker, R. J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, London: Sage.

Student (1908). The Probable Error of a Mean. *Biometrika*, 1–25.

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 paper because mixed effects models are by far the most widely used method for investigating context effects in sociology (Heisig et al., 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]It also suggests that the size of τ depends on how 'unbalanced' the clusters sizes are, that is, on how the number of lower-level observations differs across clusters. We do not pursue this point further in this paper.

[4]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 zero. 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.

[5]We have conducted additional Monte Carlo simulation results which support this claim. These results are available upon request.

[6]The simulation results indeed show that both types of models produce unbiased coefficient estimates. These results can be obtained from the replication files which are part of the online supporting material. As discussed in footnote 4, 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 .05 the true probability would be .10.

[8]Replication code for the analyses in Heisig et al. (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. (2016) 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 et al. (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]Another option might be to avoid mixed effects models altogether and use conventional regression techniques with cluster-robust variance estimation or two-step approaches (Heisig et al., 2017). However, conventional corrections for clustering are known to perform poorly with the small cluster-level samples that sociologists often deal with. More recent methods developed specifically for the few-clusters case show better performance—often involving a form of bootstrapping too—but this still is an area of active research (for details, see Cameron and Miller, 2015; Esarey and Menger, 2018).

Appendix A Standard Error Bias as an Alternative Outcome

In the main article, we focus on the actual coverage rates of two-sided 95% confidence intervals in assessing the inferential accuracy of the different estimators. An alternative approach, taken, for example, by Schmidt-Catran and Fairbrother (2015), would be to compare the average estimated standard error with the actual standard deviation of the corresponding point estimate across the Monte Carlo replications. Schmidt-Catran and Fairbrother (2015) refer to this as 'optimism of the SEs' (p.27).

However, reporting coverage has a considerable advantage. It is well known that the standard error is a downward biased estimator of the sampling distribution standard deviation when samples are small. Consider, for instance, the standard error of the mean: $\sigma(\bar{x}) = \frac{\text{SD}(x)}{\sqrt{n}}$. This estimator of the standard error relies on the sample standard deviation ($\text{SD}(x)$). Unfortunately, the latter is known to be (downward) biased estimator of the population standard deviation in small samples, even if it is based on an unbiased estimator of the population variance, as provided by the usual estimator ${\Sigma(x_i-\bar{x})^2}/{(N-1)}$ (Gurland and Tripathi, 1971). The well-established solution to this problem, going back to the work of William Gossett (1908) is to use a t-distribution with appropriate degrees of freedom for statistical inference.

Similar issues arise in the context of multilevel mixed effects regression. In particular, Elff et al. (2016) show that a t-distribution with appropriate degrees of freedom leads to accurate statistical inference for contextual (cluster-level) variables in multilevel models with few clusters. The focus on actual coverage rates in the main article allows us to implement this correction. If we focused on standard error bias, we would not be able to do this. Specifically, we would find apparent optimism of the standard errors and might misleadingly conclude that infer-

ence is anti-conservative when accurate inference is actually perfectly possible—provided that the appropriate t-distribution is used. This concern is obviously most serious for experimental conditions with few clusters.

A comparison of Tables A1 and A2 with the corresponding tables in the main article (Tables 2 and Table 3) illustrates this point. Tables A1 and A2 report relative standard error bias, that is, the difference between the average standard error estimate $\widehat{\text{SE}}(\hat{\gamma})$ and the actual standard deviation of the coefficient estimates $\text{SD}(\hat{\gamma})$ across the R Monte Carlo replications, expressed in % of $\text{SD}(\hat{\gamma})$, or formally:

$$\frac{\frac{\Sigma\widehat{\text{SE}}(\hat{\gamma})}{R} - \text{SD}(\hat{\gamma})}{\text{SD}(\hat{\gamma})} \times 100.$$

Results concerning the relative performance of the two models do not differ from the main article: standard error estimates for the cross-level interaction and the main effect of the lower-level variable generally show stronger negative bias for the model excluding the random slope than for the model including the random slope associated with the cross-level interaction. However, even the standard errors for the latter model appear to suffer from substantial negative bias, especially in the experimental conditions with only five clusters in Table A2. This contrasts very markedly with the corresponding results in the main article where we find confidence interval coverage to be largely accurate (and even slightly over-conservative in some of the more extreme experimental conditions; see Appendix B above). As discussed above, the reason for these difference is that the use of the t-distribution corrects for the substantial downward bias of the standard errors in small (cluster-level) samples.

Table A1: Standard Error Bias (%) by Variance of Lower-level Predictor and Random Slope Term

	$\gamma^{(x)}$		$\gamma^{(xz)}$		$\gamma^{(cz)}$	
	Random Slope		Random Slope		Random Slope	
SD(x_{ij})	Included	Omitted	Included	Omitted	Included	Omitted
	$R^2(\beta_j^{(x)}) = 0.95$ (i.e., SD($u_j^{(x)}$) ≈ 0.23)					
0.5	0.76	-16.72	-0.77	-17.98	-3.91	-3.88
1.0	-2.39	-39.03	-3.67	-39.89	-5.27	-5.06
2.0	-1.44	-63.37	-4.32	-64.42	-3.79	-3.42
	$R^2(\beta_j^{(x)}) = 0.90$ (i.e., SD($u_j^{(x)}$) ≈ 0.33)					
0.5	-0.74	-27.12	-2.65	-28.44	-3.64	-3.56
1.0	-1.43	-52.43	-4.67	-53.94	-2.52	-2.19
2.0	-2.02	-73.80	-3.38	-74.17	-3.74	-3.23
	$R^2(\beta_j^{(x)}) = 0.50$ (i.e., SD($u_j^{(x)}$) $= 1.00$)					
0.5	-1.39	-65.97	-2.69	-66.42	-4.32	-4.14
1.0	-2.46	-81.97	-3.77	-82.20	-4.57	-4.21
2.0	-1.11	-90.10	-4.83	-90.47	-4.48	-4.01
	$R^2(\beta_j^{(x)}) = 0.10$ (i.e., SD($u_j^{(x)}$) $= 3.00$)					
0.5	-2.81	-87.57	-4.43	-87.77	-4.68	-4.47
1.0	-0.96	-92.71	-3.41	-92.88	-3.74	-3.15
2.0	-1.83	-94.92	-2.27	-94.95	-5.06	-3.74

Note: Results are based on 10,000 Monte Carlo replications. Note that for reasons of brevity, this table does not express Monte Carlo error. The number of observations per cluster is 500 with overall 15 clusters.

Table A2: Standard Error Bias (%) by Number of Clusters and Lower-level Observations

		$\gamma^{(x)}$		$\gamma^{(xz)}$		$\gamma^{(cz)}$	
		Random Slope		Random Slope		Random Slope	
n_j	n_{total}	Included	Omitted	Included	Omitted	Included	Omitted
		$m = 5$ Clusters					
100	500	-8.93	-63.16	-22.94	-68.69	-16.97	-16.24
500	2500	-9.82	-82.54	-15.12	-83.55	-19.93	-18.65
1000	5000	-13.48	-87.93	-21.46	-89.08	-19.03	-18.36
		$m = 15$ Clusters					
100	1500	-2.26	-62.20	-4.71	-63.04	-7.90	-3.96
500	7500	-2.46	-81.97	-3.77	-82.20	-4.57	-4.21
1000	15000	-2.30	-87.16	-3.81	-87.36	-3.79	-3.72
		$m = 25$ Clusters					
100	2500	-1.73	-62.30	-2.95	-62.69	-5.02	-1.90
500	12500	-1.00	-81.92	-2.33	-82.16	-1.36	-1.33
1000	25000	-1.19	-87.12	-1.66	-87.19	-1.68	-1.68

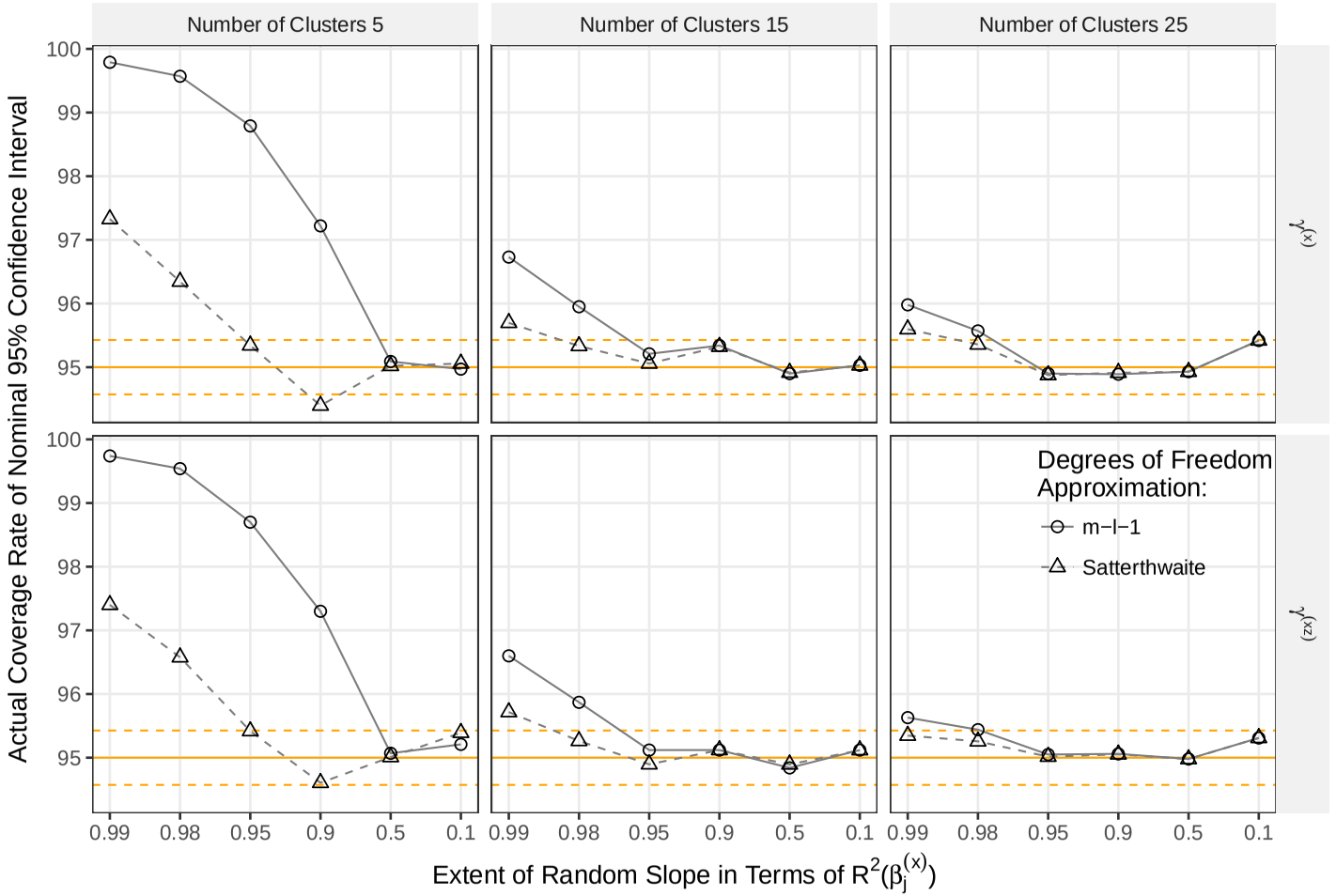
Note: Results are based on 10,000 Monte Carlo replications. Note that for reasons of brevity, this table does not express Monte Carlo error. These results are based on experimental conditions for which $R^2(\beta_j^{(x)}) = 0.50$ (i.e., SD($u_j^{(x)}$) $= 1$), and SD(x_{ij}) $= 1$.

Appendix B At the Limit: When $R^2(\beta_j^{(x)})$ is Large and the Cluster Sample Small

The simulation results in the main article clearly show that models with cross-level interactions should generally include a random slope on the corresponding lower-level components. However, Tables 2 and 3 in the main article also suggest that such models may produce overconservative inference in extreme situations when a) the number of clusters is very small ($m = 5$ in our simulations) or when b) the random slope exhibits very little unexplained variability (SD($u_j^{(x)}$) ≈ 0.33, corresponding to an upper-level $R^2(\beta_j^{(x)})$ of 0.95). In these situations the actual coverage rates of two-sided 95% confidence intervals exceed their nominal level. With respect to significance testing, this means that a true null hypothesis will be rejected less frequently than the nominal level of the test suggests.

Do these results warrant a qualification of the recommendation to always include a random slope on the lower-level component of a cross-level interaction? We would argue that the answer is almost always *no* because overcoverage only arises under extreme conditions that have little practical relevance. This reassuring result notwithstanding, this appendix presents additional analyses that reduce the variability of the random slope even further, pushing $R^2(\beta_j^{(x)})$ beyond 0.95 and very close to 1. These are situations where the error in the upper-level model for $\beta_j^{(x)}$ exhibits very little variation, so there remains very little 'clustering' in the sense of correlated errors for lower-level units belonging to the same cluster. At least, that is, to the extent that such correlation is due to unobserved cluster-specific differences in the relationship between y_{ij} and x_{ij}; there may still be cluster-correlated errors due to a random intercept term or to random slope terms on other lower-level variables. When clustering becomes negligible in this way, the $m - l - 1$ rule for approximating the degrees of freedom for confidence intervals and t-tests may no longer work well because it is based on the idea

Figure B1: Statistical Inference for a Cross-Level Interaction Term at the Limits



Note: These results are based on 500 observations per cluster and the standard deviation of the lower-level predictor is set to 1.

that $m - l - 1$ would be the correct degrees of freedom in the implicit cluster-/aggregate-level regression (Elff et al., 2016). Therefore, we also consider an alternative, computationally more intensive approximation, a generalization of the Satterthwaite (1946) method that was first proposed by Giesbrecht and Burns (1985; for an overview of degree of freedom approximations in the mixed effects context, see Schaalje et al. 2002). Elff et al. (2016) find the Satterthwaite method to perform very similarly to the $m - l - 1$ rule, but they do not consider the kinds of extreme situations where the above analysis shows the latter approach to produce overconservative inference.[13]

Figure B1 plots the actual coverage rates of confidence intervals for the cross-level interaction term and the main effect of its lower-level component. Solid lines show coverage rates for confidence intervals based on the $m - l - 1$ rule; dashed lines show coverage rates for intervals based on the Satterthwaite approximation. Whereas the most extreme case considered so far was that of an implied upper-level $R^2(\beta_j^{(x)})$ of 0.95, we now consider two additional cases with $R^2(\beta_j^{(x)})$ values of .98 and .99, respectively. In these situations, there is almost no unexplained cross-cluster variation in $\beta_j^{(x)}$ and arguably much less than one could expect to encounter in most social science applications.

Figure B1 confirms the one qualification of our recommendation to always include a random slope on the lower-level components of cross-level interaction terms: in cases where the variance of the random slope term is extremely small, following this recommendation can result in overconservative inference, especially if the number of clusters is also very low. The problems seems to at least partly stem from the inaccuracy of the $m - l - 1$ approximation to the degrees of freedom, as confidence intervals based on the Satterthwaite method perform much better under extreme conditions. However, even the Satterthwaite method fails in the most extreme scenarios.

One might alternatively suspect that convergence problems are responsible

for the overcoverage because estimation of a near-zero variance component can create problems for the optimization process. Yet, disaggregated analysis of Monte Carlo trials with and without convergence warnings provides no support for this explanation. These results can be obtained from the replication files which are part of the online supporting material.

While the findings of this section warrant a note of caution, we would like to emphasize again that both methods yield accurate statistical inference under practically relevant conditions (15 or more clusters and $R^2(\beta_j^{(x)}) \leq 0.9$), whereas the results presented in the main article show models that omit the crucial random slope term to produce overly optimistic results in such situations.

Appendix C Model Selection Criteria are no Remedy

The simulation results in the main article suggest that practitioners who analyze cross-level interactions using mixed effects models are well-advised to always include a random slope on the lower-level component. However, instead of opting for a random slope on *a priori* grounds, one might take a more data-driven approach and rely on standard model selection criteria such as likelihood ratio tests or information measures such as AIC and BIC in choosing a random effects specification. For example, 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', established model selection criteria are obvious candidates when it comes to determining whether a given slope is small enough to warrant omission.

Perhaps unsurprisingly, we would not recommend to rely on model selection criteria in determining whether to include the random slope associated with a cross-level interaction. For reasons given in Section 'Why *Always* a Random Slope?' in the main article, we would argue these random slopes should *always* be included in all practically relevant situations. To support this claim, this appendix summarizes additional Monte Carlo evidence demonstrating that data-driven approaches based on model selection criteria will lead to substantial undercoverage in at least some situations. We investigate this issue as follows: for each simulated data set, we determine whether a given model selection criterion favors the model with or the model without the random slope on the lower-level component. To assess the performance of a given selection criterion, we then calculate the actual coverage rate of the models thus selected.

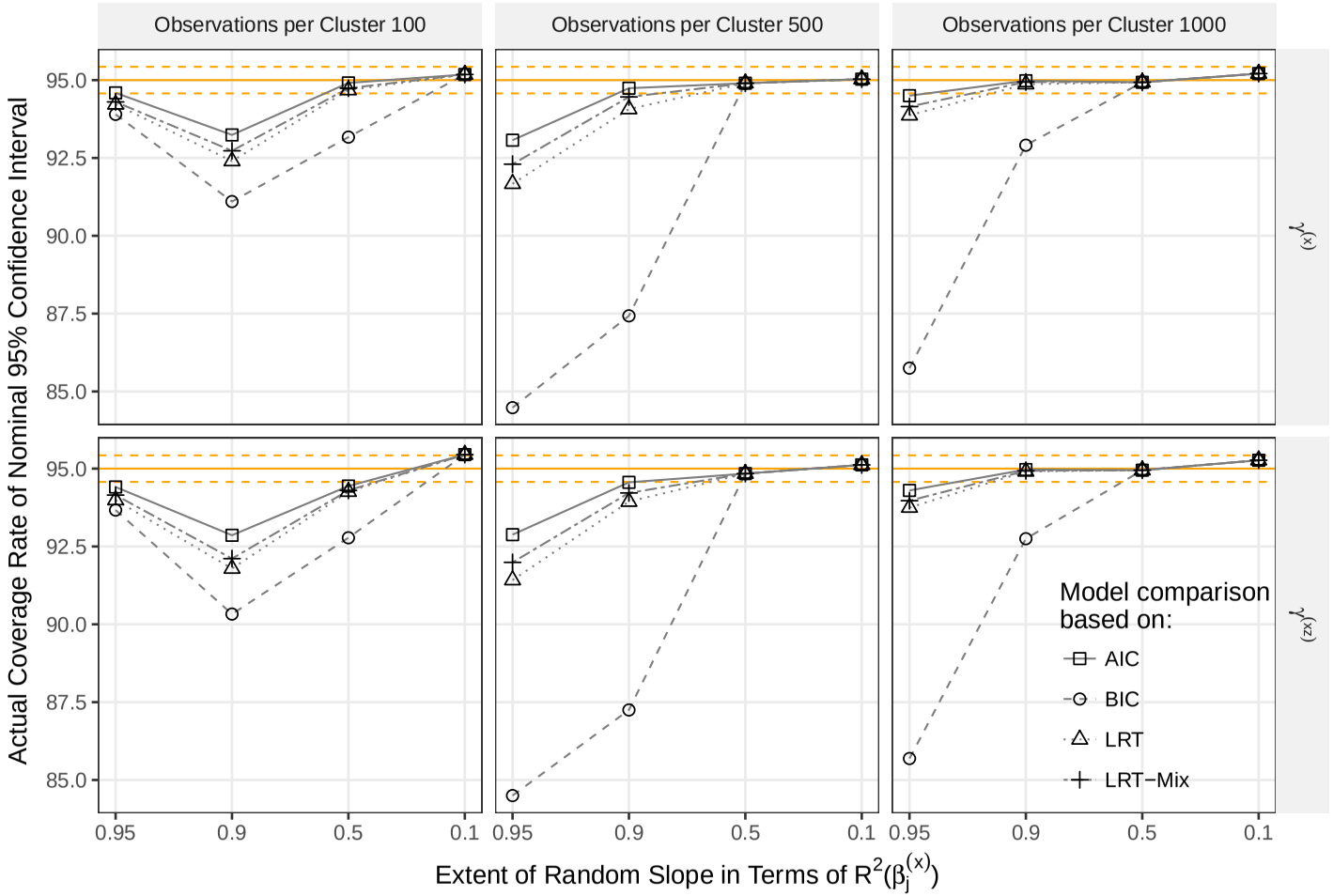
We consider four model selection criteria. The first two are variants of a likelihood ratio/deviance test (LRT). Both are based on the difference in the deviance

statistic (i.e., -2 times the log likelihood) between the random intercept model and the model that includes the random slope term as well as its covariance with the random intercept. The first variant compares the difference in the deviance against a Chi-Square distribution with two degrees of freedom (one for the slope variance and one for the covariance). The null hypothesis of the test is that the variance and covariance parameter are jointly zero, so we choose the model including the random slope when the test result is significant ($p < .05$) and the random intercept model otherwise. It is well-known that this test is overconservative (i.e., underrejects the null hypothesis) because the variance parameter cannot be smaller than zero. The second variant therefore uses the average of the p-values obtained from Chi-square distributions with one and two degrees of freedom (see, for example, Snijders and Bosker, 2012, 98f.). In addition to the two variants of the LRT, we consider Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) as alternative selection criteria. We used *R*'s *anova* function to calculate the deviance statistics and information criteria, which uses the likelihood from maximum likelihood rather than restricted maximum likelihood estimation. Confidence intervals for the calculation of coverage rates are based on restricted maximum likelihood estimates, however.

Figure C1 plots the actual coverage rates of the confidence intervals for the cross-level interaction term and the main effect of the lower-level component because only these are affected by omitting the random slope term (see Table 2 in the main article). We focus on a subset of the experimental conditions. In particular, we show results for 15 clusters and a standard deviation of 1 on the lower-level predictor. Results for the other experimental conditions do not lead to qualitatively different conclusions and can be obtained from the replication files which are part of the online supporting material.

The overall message emerging from Figure C1 is clear: when the goal is to achieve correct statistical inference for a cross-level interaction effect, it is not

Figure C1: Actual Coverage Rates for Different Model Selection Strategies



Note: These results are based on 15 clusters; the standard deviation of the lower-level predictor is set to 1.

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. This is not surprising, as we saw above that models that include the random slope term on the lower-level component generally lead to accurate inference, whereas models that omit the term suffer from undercoverage—with the extent of undercoverage depending on various aspects of the DGP. The model selection criteria investigated here will sometimes favor the model including the random slope, and sometimes the one omitting it. The coverage rate for a given model selection strategy will thus be a weighted average of the coverage rates for the correct model (i.e., the one with a random intercept and slope) and for the misspecified model (i.e., the one with only a random intercept). Thus, taking a data-driven approach to model selection will generally be better than selecting the model without a random slope *a priori*, but only because it sometimes favors the model including the random slope.

Detailed inspection of Figure C1 reveals some interesting patterns. The first is that model selection based on BIC performs worst and model selection based on AIC best, with the two variants of the LRT falling in between. LRTs using a mixture of Chi-Square distributions with one and two degrees of freedom have a slight edge over the alternative because they more often reject the random intercept model. The reason why BIC performs more poorly than the other criteria is that it penalizes additional parameters more harshly, particularly in large samples, so it more often favors the model omitting the random slope, which is more parsimonious (BIC uses a penalty of $\log(n)$, whereas AIC uses a constant penalty of 2; Burnham and Anderson, 2004). Another noteworthy pattern is that the performance of all four model selection strategies improves as the (implicit) $R^2(\beta_j^{(x)})$ of the cluster-level regression for the slope of x_{ij} declines or, equivalently, as the standard deviation of the random slope (i.e., SD($u_j^{(x)}$)) in the DGP increases. Intu-

itively, this is because all model selection strategies become more likely to favor the model that includes the random slope, the more variation the latter shows. Finally, the performance of the different model selection strategies depends on the number of observations per cluster. Model selection based on AIC and the two variants of the LRT tends to improve as the number of observations per cluster increases (except when the random slope shows very little variation with an implied cluster-level $R^2(\beta_j^{(x)})$ of 0.95). This is because both the LRT and AIC more often favor the model that includes the random slope in larger samples. The reason why BIC performs differently from AIC again is that it penalizes additional parameters using a factor that depends on the sample size.

Overall, the impact of the cluster-level $R^2(\beta_j^{(x)})$ and the lower-level sample size on the performance of the different model selection strategies should be taken as illustrative. Their performance in applied settings will depend on various other (and partly unobservable) aspects of a given analysis. The main message to take away from Figure C1 is that there are practically relevant situations where reliance on model selection criteria will lead to anticonservative inference for the cross-level interaction. These results make very clear that one should not blindly rely on model selection criteria in determining whether to include a random slope on the lower-level component of a cross-level interaction. Rather, as we emphasize in the main article, the default should be to specify a random slope term, so much so that we would practically recommend to *always* include it. There may be a very limited role for model selection criteria in situations characterized by negligible slope variaton (see Appendix B above), but the results presented in this section show that selection criteria must not be the only factor taken into account, as they can easily lead to severely anti-conservative inference (in particular, the substantive magnitude of cross-cluster variation in the slope should be considered as well). Moreover, as also emphasized in the main article, we believe that situations where variation is so low that omitting the random

slope might be a reasonable choice are rare exceptions in practice, at least for typical sociological application. Our empirical examples (see Appendix D below) where we generally find substantive variation in the random slopes even after including the cross-level interactions with HDI support this view (see the final columns of Tables D1 to D6 below). However, while we strongly suspect that these findings generalize to most other applications, we do not hesitate to admit that this is ultimately an empirical question that we cannot answer within the confines of our study.

Appendix D Additional Illustrative Empirical Analyses

To get a sense of how serious the consequences of omitting the random slope term for a cross-level interaction are in real-world applications, we conduct a series of illustrative analyses based on European Social Survey data (ESS Round 6, 2016). We adopt Heisig et al.'s (2017) illustrative analyses of cross-level interactions. Replication code for the analyses in Heisig et al. (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.

Our 30 empirical examples study how the relationships between six lower-level predictors (having a high education, age, gender, unemployment, being married, and having a medium education) and five standard outcome variables (generalized trust, homophobia, xenophobia, fear of crime, and occupational status) are moderated by the Human Development Index (HDI). For each of the 30 illustrative cross-level interactions we estimate a specification including and one omitting the random slope term for the lower-level variable involved in the respective cross-level interaction. Overall, this results in 60 linear mixed effects models.

All outcome variables and age are standardized to have a mean of zero and a standard deviation of one. Education is measured as an individual's highest degree, subsumed into three categories: low (highest degree below the upper secondary level), intermediate (highest degree at the upper secondary or non-tertiary, post-secondary level), and high (highest degree at the tertiary level). Being female, being married, and being unemployed (ILO definition) are also indicator variables. Following Heisig et al. (2017), all indicator variables are weighted effects (rather than dummy) coded. Weighted effects coding of categorical pre-

dictors is akin to grand mean centering of continuous predictors and ensures that the intercept corresponds to the predicted outcome for the 'average' individual (Grotenhuis et al., 2016). The coefficient of the high education indicator, for instance, captures the (adjusted) difference in the respective outcome variable (e.g., fear of crime) between high-educated individuals and individuals whose level of education equals that of the average European. Its cross-level interaction with the HDI indicates whether this difference changes with a society's level of human development. Due to the presence of the cross-level interaction term the main effect of the high education indicator must be interpreted as the conditional effect of having high education for a country with an HDI of zero, that is, for a country with an average level of human development. 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. These controls are not interacted with other (lower- or upper-level) predictors. Further details are described in Heisig et al. (2017).

Tables D1 to D6 present a summary of the main results, omitting coefficient estimates for control variables. Results for fear of crime at the top of Table D1 show that the cross-level interaction between the HDI and having high education is negative and statistically significant, irrespective of whether we include a random slope term or not. The same holds for the main effect of being highly educated. Thus, the high educated tend to be less afraid of crime than Europeans with average education and their advantage in (perceived) security is particularly strong in countries with a high degree of human development.

Qualitatively, this conclusion does not depend on the random effects specification, but the model that does not contain a random slope for high education strongly overstates the precision with which we can estimate the cross-level interaction and the main effect of high education. The third column shows that the estimated standard errors (in parentheses) for these coefficients are substan-

tially larger in the correctly specified model that includes the random slope—by 67.4% for the main effect and by 82.3% for the cross-level interaction. Accordingly, the absolute t-ratios (in brackets) are much smaller when the model is correctly specified—by 40.8% for the main effect and by 46.8% for the cross-level interaction. Over the 30 different models (5 dependent variables $\times$ 6 lower-level predictors), the reduction in the absolute t-ratio for the cross-level interaction effect due to including the random slope is 42.4% on average. The median reduction is 48.3% and the 25th and 75th percentiles are 31.3 and 60.9%, respectively. Figure 2 provides a compact visual representation of the results. For all 30 cross-level interactions, it shows how the t-ratio of the interaction term changes due to the inclusion of associated random slope.

Table D1: Cross-Level Interaction of High Education and the HDI for Five Outcomes

	Random Slope		Δ	
	Included	Omitted	in %	$\frac{\mathrm{SD}(u_j^{(x)})}{\beta^{(x)}}$
		Fear of Crime		
High education	-0.107^{***}	-0.108^{***}		
	(0.014)	(0.008)	67.429	
	[7.518]	[12.693]	−40.775	
HDI	-0.256^{***}	-0.257^{***}		
	(0.042)	(0.042)	0.316	
	[6.087]	[6.134]	−0.771	
HDI*High education	-0.071^{***}	-0.074^{***}		
	(0.014)	(0.008)	82.263	
	[5.137]	[9.657]	−46.799	0.556
		Generalized Trust		
High education	0.209^{***}	0.203^{***}		
	(0.016)	(0.008)	88.940	
	[13.323]	[24.501]	−45.624	
HDI	0.347^{***}	0.349^{***}		
	(0.060)	(0.059)	0.638	
	[5.796]	[5.874]	−1.319	
HDI*High education	0.038^{*}	0.033^{***}		
	(0.015)	(0.007)	107.557	
	[2.453]	[4.451]	−44.894	0.334
		Homophobia		
High education	-0.170^{***}	-0.160^{***}		
	(0.019)	(0.008)	143.331	
	[8.966]	[20.474]	−56.206	
HDI	-0.453^{***}	-0.456^{***}		
	(0.065)	(0.065)	0.366	
	[6.995]	[7.061]	−0.932	
HDI*High education	−0.005	−0.002		
	(0.019)	(0.007)	170.710	
	[0.280]	[0.331]	−15.401	0.534
		Occupational Status (ISEI)		
High education	1.033^{***}	1.028^{***}		
	(0.013)	(0.007)	80.285	
	[78.859]	[141.441]	−44.246	
HDI	0.110^{***}	0.114^{***}		
	(0.024)	(0.023)	0.769	
	[4.682]	[4.857]	−3.598	
HDI*High education	-0.047^{**}	-0.051^{***}		
	(0.013)	(0.007)	97.488	
	[3.667]	[7.883]	−53.480	0.055
		Xenophobia		
High education	-0.320^{***}	-0.314^{***}		
	(0.021)	(0.008)	162.675	
	[15.403]	[39.736]	−61.237	
HDI	-0.126^{+}	-0.132^{+}		
	(0.069)	(0.070)	−1.053	
	[1.819]	[1.880]	−3.279	
HDI*High education	-0.072^{**}	-0.071^{***}		
	(0.021)	(0.007)	193.021	
	[3.442]	[10.022]	−65.652	0.316

Note: Estimates are from linear mixed effects models. All estimates are controlled for: age, marital status, unemployment, intermediate, and high (compared to low) education. Standard errors in parentheses, absolute t-statistics in brackets. $^{+}p < 0.1$;$^{*}p < 0.05$;$^{**}p < 0.01$;$^{***}p < 0.001$. The p-values for HDI and in the models including a random slope also the p-values for high education are based on the t-distribution with degrees of freedom approximated by the $m - l - 1$ rule (c.f., Elff et al., 2016). p-value for high education in the model without a random slope is based on the normal distribution.

Table D2: Cross-Level Interaction of Gender and the HDI for Five Outcomes

	Random Slope		Δ	
	Included	Omitted	in %	$\frac{\mathrm{SD}(u_j^{(x)})}{\beta^{(x)}}$
		Fear of Crime		
Female	0.209***	0.209***		
	(0.013)	(0.005)	198.409	
	[15.465]	[46.242]	−66.556	
HDI	−0.259***	−0.260***		
	(0.043)	(0.043)	−0.109	
	[6.025]	[6.030]	−0.087	
HDI*Female	0.029*	0.034***		
	(0.014)	(0.005)	185.902	
	[2.112]	[6.987]	−69.768	0.321
		Generalized Trust		
Female	0.020**	0.021***		
	(0.007)	(0.004)	50.947	
	[3.054]	[4.780]	−36.115	
HDI	0.351***	0.350***		
	(0.059)	(0.059)	−0.0003	
	[5.928]	[5.919]	0.150	
HDI*Female	0.005	0.005		
	(0.007)	(0.005)	47.118	
	[0.660]	[1.019]	−35.200	1.285
		Homophobia		
Female	−0.084***	−0.085***		
	(0.007)	(0.004)	68.905	
	[12.061]	[20.462]	−41.058	
HDI	−0.455***	−0.456***		
	(0.065)	(0.065)	0.029	
	[7.052]	[7.062]	−0.140	
HDI*Female	−0.012+	−0.013**		
	(0.007)	(0.004)	64.039	
	[1.712]	[2.951]	−41.974	0.349
		Occupational Status (ISEI)		
Female	0.009	0.011**		
	(0.010)	(0.004)	157.683	
	[0.855]	[2.879]	−70.293	
HDI	0.114***	0.112***		
	(0.023)	(0.023)	−0.581	
	[5.010]	[4.912]	1.996	
HDI*Female	−0.015	−0.015**		
	(0.010)	(0.004)	147.440	
	[1.426]	[3.669]	−61.142	5.666
		Xenophobia		
Female	0.002	0.002		
	(0.008)	(0.004)	93.748	
	[0.239]	[0.512]	−53.223	
HDI	−0.136+	−0.134+		
	(0.070)	(0.070)	−0.263	
	[1.937]	[1.906]	1.632	
HDI*Female	−0.004	−0.003		
	(0.008)	(0.005)	87.324	
	[0.488]	[0.579]	−15.741	18.702

Note: Estimates are from linear mixed effects models. All estimates are controlled for: age, marital status, unemployment, intermediate, and high (compared to low) education. Standard errors in parentheses, absolute t-statistics in brackets. $^{+}p < 0.1;^{*}p < 0.05;^{**}p < 0.01;^{***}p < 0.001$. The p-values for HDI and in the models including a random slope also the p-values for high education are based on the t-distribution with degrees of freedom approximated by the $m - l - 1$ rule (c.f., Elff et al., 2016). p-value for high education in the model without a random slope is based on the normal distribution.

Table D3: Cross-Level Interaction of Age and the HDI for Five Outcomes

	Random Slope		Δ	
	Included	Omitted	in %	$\frac{\mathrm{SD}(u_j^{(x)})}{\beta^{(x)}}$
		Fear of Crime		
Age	0.070^{***}	0.072^{***}		
	(0.012)	(0.005)	134.255	
	[5.925]	[14.264]	−58.460	
HDI	-0.259^{***}	-0.259^{***}		
	(0.043)	(0.043)	0.290	
	[6.035]	[6.049]	−0.234	
HDI*Age	0.009	0.006		
	(0.012)	(0.005)	134.058	
	[0.787]	[1.124]	−30.026	0.800
		Generalized Trust		
Age	0.027^{*}	0.030^{***}		
	(0.011)	(0.005)	132.105	
	[2.379]	[6.030]	−60.543	
HDI	0.351^{***}	0.351^{***}		
	(0.059)	(0.059)	0.137	
	[5.923]	[5.939]	−0.269	
HDI*Age	0.022^{+}	0.022^{***}		
	(0.012)	(0.005)	131.911	
	[1.864]	[4.310]	−56.755	1.987
		Homophobia		
Age	0.141^{***}	0.141^{***}		
	(0.013)	(0.005)	184.520	
	[10.705]	[30.549]	−64.960	
HDI	-0.456^{***}	-0.457^{***}		
	(0.064)	(0.065)	−0.585	
	[7.088]	[7.053]	0.488	
HDI*Age	-0.032^{*}	-0.034^{***}		
	(0.013)	(0.005)	184.332	
	[2.424]	[7.229]	−66.468	0.460
		Occupational Status (ISEI)		
Age	0.083^{***}	0.082^{***}		
	(0.010)	(0.004)	122.890	
	[8.601]	[18.987]	−54.701	
HDI	0.111^{***}	0.112^{***}		
	(0.023)	(0.023)	0.400	
	[4.881]	[4.946]	−1.299	
HDI*Age	0.003	0.005		
	(0.010)	(0.004)	122.678	
	[0.265]	[1.027]	−74.200	0.544
		Xenophobia		
Age	0.087^{***}	0.088^{***}		
	(0.014)	(0.005)	189.628	
	[6.428]	[18.767]	−65.747	
HDI	-0.135^{+}	-0.135^{+}		
	(0.069)	(0.070)	−1.507	
	[1.943]	[1.911]	1.656	
HDI*Age	−0.011	-0.013^{*}		
	(0.014)	(0.005)	189.452	
	[0.816]	[2.715]	−69.936	0.768

Note: Estimates are from linear mixed effects models. All estimates are controlled for: age, marital status, unemployment, intermediate, and high (compared to low) education. Standard errors in parentheses, absolute t-statistics in brackets. $^{+}p < 0.1;^{*}p < 0.05;^{**}p < 0.01;^{***}p < 0.001$. The p-values for HDI and in the models including a random slope also the p-values for high education are based on the t-distribution with degrees of freedom approximated by the $m - l - 1$ rule (c.f., Elff et al., 2016). p-value for high education in the model without a random slope is based on the normal distribution.

Table D4: Cross-Level Interaction of Marital Status and the HDI for Five Outcomes

	Random Slope		Δ	
	Included	Omitted	in %	$\frac{\mathrm{SD}(u_j^{(x)})}{\beta^{(x)}}$
		Fear of Crime		
Married	-0.018^{*}	-0.019^{***}		
	(0.008)	(0.004)	68.128	
	[2.443]	[4.247]	−42.479	
HDI	-0.260^{***}	-0.259^{***}		
	(0.043)	(0.043)	−0.573	
	[6.098]	[6.029]	1.132	
HDI*Married	−0.004	−0.007		
	(0.008)	(0.005)	65.587	
	[0.566]	[1.585]	−64.286	1.713
		Generalized Trust		
Married	0.022^{***}	0.023^{***}		
	(0.005)	(0.004)	13.491	
	[4.533]	[5.177]	−12.434	
HDI	0.350^{***}	0.350^{***}		
	(0.059)	(0.059)	−0.082	
	[5.915]	[5.913]	0.037	
HDI*Married	0.013^{*}	0.013^{**}		
	(0.005)	(0.005)	12.963	
	[2.484]	[2.787]	−10.900	0.538
		Homophobia		
Married	0.016^{**}	0.016^{***}		
	(0.005)	(0.004)	23.272	
	[3.221]	[3.991]	−19.288	
HDI	-0.456^{***}	-0.456^{***}		
	(0.065)	(0.065)	0.024	
	[7.060]	[7.062]	−0.018	
HDI*Married	0.001	0.002		
	(0.005)	(0.004)	22.412	
	[0.176]	[0.358]	−50.913	0.936
		Occupational Status (ISEI)		
Married	0.027^{***}	0.026^{***}		
	(0.006)	(0.004)	53.269	
	[4.552]	[6.892]	−33.954	
HDI	0.112^{***}	0.112^{***}		
	(0.023)	(0.023)	0.085	
	[4.927]	[4.914]	0.278	
HDI*Married	0.004	0.004		
	(0.006)	(0.004)	51.331	
	[0.604]	[0.938]	−35.540	0.863
		Xenophobia		
Married	0.002	0.001		
	(0.006)	(0.004)	35.386	
	[0.277]	[0.348]	−20.481	
HDI	-0.133^{+}	-0.134^{+}		
	(0.070)	(0.070)	−0.016	
	[1.893]	[1.898]	−0.283	
HDI*Married	−0.007	−0.007		
	(0.006)	(0.004)	34.175	
	[1.133]	[1.527]	−25.793	12.559

Note: Estimates are from linear mixed effects models. All estimates are controlled for: age, marital status, unemployment, intermediate, and high (compared to low) education. Standard errors in parentheses, absolute t-statistics in brackets. $^{+}p < 0.1;^{*}p < 0.05;^{**}p < 0.01;^{***}p < 0.001$. The p-values for HDI and in the models including a random slope also the p-values for high education are based on the t-distribution with degrees of freedom approximated by the $m - l - 1$ rule (c.f., Elff et al., 2016). p-value for high education in the model without a random slope is based on the normal distribution.

Table D5: Cross-Level Interaction of Being Unemployed and the HDI for Five Outcomes

	Random Slope		Δ	
	Included	Omitted	in %	$\frac{\mathrm{SD}(u_j^{(x)})}{\beta^{(x)}}$
		Fear of Crime		
Unemployed	0.064*	0.052**		
	(0.025)	(0.019)	32.738	
	[2.531]	[2.700]	−6.262	
HDI	−0.259***	−0.259***		
	(0.043)	(0.043)	0.078	
	[6.031]	[6.050]	−0.311	
HDI*Unemployed	0.002	−0.005		
	(0.026)	(0.020)	32.217	
	[0.069]	[0.247]	−72.198	1.265
		Generalized Trust		
Unemployed	−0.133***	−0.135***		
	(0.019)	(0.019)	0.416	
	[7.105]	[7.192]	−1.206	
HDI	0.351***	0.351***		
	(0.059)	(0.059)	0.019	
	[5.925]	[5.926]	−0.013	
HDI*Unemployed	−0.024	−0.023		
	(0.020)	(0.019)	0.417	
	[1.220]	[1.208]	0.948	0.064
		Homophobia		
Unemployed	0.025	0.025		
	(0.018)	(0.018)	0.817	
	[1.399]	[1.398]	0.103	
HDI	−0.456***	−0.456***		
	(0.064)	(0.064)	0.001	
	[7.077]	[7.077]	−0.001	
HDI*Unemployed	0.040*	0.040*		
	(0.018)	(0.018)	0.802	
	[2.183]	[2.214]	−1.364	0.415
		Occupational Status (ISEI)		
Unemployed	−0.215***	−0.206***		
	(0.023)	(0.016)	39.448	
	[9.400]	[12.582]	−25.290	
HDI	0.111***	0.112***		
	(0.023)	(0.023)	0.299	
	[4.900]	[4.943]	−0.887	
HDI*Unemployed	−0.011	−0.006		
	(0.024)	(0.017)	38.842	
	[0.462]	[0.345]	33.847	0.364
		Xenophobia		
Unemployed	0.077**	0.080***		
	(0.025)	(0.018)	39.156	
	[3.085]	[4.504]	−31.516	
HDI	−0.135+	−0.135+		
	(0.070)	(0.070)	−0.206	
	[1.923]	[1.915]	0.396	
HDI*Unemployed	0.027	0.033+		
	(0.026)	(0.019)	38.488	
	[1.041]	[1.756]	−40.694	1.112

Note: Estimates are from linear mixed effects models. All estimates are controlled for: age, marital status, unemployment, intermediate, and high (compared to low) education. Standard errors in parentheses, absolute t-statistics in brackets. $^{+}p < 0.1;^{*}p < 0.05;^{**}p < 0.01;^{***}p < 0.001$. The p-values for HDI and in the models including a random slope also the p-values for high education are based on the t-distribution with degrees of freedom approximated by the $m - l - 1$ rule (c.f., Elff et al., 2016). p-value for high education in the model without a random slope is based on the normal distribution.

Table D6: Cross-Level Interaction of Intermediate Education and the HDI for Five Outcomes

	Random Slope		Δ	
	Included	Omitted	in %	$\frac{\mathrm{SD}(u_j^{(x)})}{\beta^{(x)}}$
		Fear of Crime		
Intermediate education	0.011	0.010*		
	(0.007)	(0.005)	53.577	
	[1.540]	[2.012]	−23.440	
HDI	−0.261***	−0.259***		
	(0.043)	(0.042)	0.187	
	[6.146]	[6.102]	0.714	
HDI*Intermediate education	−0.014+	−0.018***		
	(0.007)	(0.004)	64.316	
	[1.902]	[4.202]	−54.731	2.560
		Generalized Trust		
Intermediate education	−0.017+	−0.019***		
	(0.008)	(0.005)	79.647	
	[2.020]	[4.147]	−51.285	
HDI	0.351***	0.350***		
	(0.060)	(0.059)	0.314	
	[5.889]	[5.895]	−0.114	
HDI*Intermediate education	0.010	0.010*		
	(0.008)	(0.004)	94.332	
	[1.188]	[2.368]	−49.827	2.148
		Homophobia		
Intermediate education	0.008	0.009*		
	(0.006)	(0.004)	36.273	
	[1.378]	[1.985]	−30.562	
HDI	−0.455***	−0.456***		
	(0.065)	(0.065)	0.215	
	[7.018]	[7.051]	−0.456	
HDI*Intermediate education	0.004	0.005		
	(0.006)	(0.004)	44.062	
	[0.731]	[1.270]	−42.450	2.546
		Occupational Status (ISEI)		
Intermediate education	−0.135***	−0.137***		
	(0.007)	(0.004)	74.316	
	[19.081]	[33.593]	−43.201	
HDI	0.110***	0.112***		
	(0.023)	(0.023)	1.505	
	[4.700]	[4.875]	−3.573	
HDI*Intermediate education	−0.010	−0.014***		
	(0.007)	(0.004)	88.203	
	[1.473]	[3.714]	−60.327	0.224
		Xenophobia		
Intermediate education	0.037***	0.038***		
	(0.009)	(0.004)	98.740	
	[4.222]	[8.561]	−50.687	
HDI	−0.134+	−0.134+		
	(0.071)	(0.070)	0.395	
	[1.894]	[1.905]	−0.610	
HDI*Intermediate education	−0.005	−0.005		
	(0.009)	(0.004)	116.128	
	[0.619]	[1.258]	−50.842	1.071

Note: Estimates are from linear mixed effects models. All estimates are controlled for: age, marital status, unemployment, intermediate, and high (compared to low) education. Standard errors in parentheses, absolute t-statistics in brackets. $^{+}p < 0.1;^{*}p < 0.05;^{**}p < 0.01;^{***}p < 0.001$. The p-values for HDI and in the models including a random slope also the p-values for high education are based on the t-distribution with degrees of freedom approximated by the $m - l - 1$ rule (c.f., Elff et al., 2016). p-value for high education in the model without a random slope is based on the normal distribution.

Appendix E P-Curve Analysis

In this section, we provide a more detailed analysis of the possibility that cross-level interaction estimates published in the ESR are subject to selective reporting due to publication bias and/or p-hacking. By publication bias we mean a tendency that statistically significant results with $p < .05$ are more likely to be published than 'null results' with $p \geq .05$. Publication bias could arise because editors and referees have a preference for publishing significant results. The findings of Franco et al. (2014), however, suggest that the primary reason for publication bias is that authors do not even submit insignificant results for publication, potentially because they anticipate that chances of eventual acceptance are slim. By p-hacking we mean that researchers may (consciously or unconsciously) engage in behaviors that 'push' p below .05. For example, a researcher might decide to collect additional data when findings are not (yet) significant or he/she might change regression specifications in order to obtain significant results. Both publication bias and p-hacking can artificially inflate the apparent strength of empirical support for a hypothesis.

Our analysis draws on work by Simonsohn et al. (2014, 2015), who propose p-curve analysis as a method for detecting publication bias and p-hacking on the aggregate level. The Simonsohn et al. (2014) article gives a very good overview, which is why we only give a brief summary of the approach. The p-curve approach circumvents the problem that insignificant results remain unpublished by assessing the evidential value of a collection of studies on the basis of statistically significant (published) results only. The p-curve describes the relative frequency of different p-values below the .05 threshold. On the aggregate level, a collection of studies that has evidential value (i.e., that at least partly reports results on effects or associations that really exist) will produce a right-skewed distribution. That is, smaller p-values should be more likely to occur than higher ones. In other words, 'highly significant' results with, say, $p < .01$ should be observed

more often than 'just-significant' results with a p-value of, say, .49. By contrast, if an effect does not really exist (i.e., if the null-hypothesis is correct), the p-curve will be uniform. A uniform p-curve hence indicates publication bias: the published significant studies lack evidential basis. The fact that there seems to be positive empirical support for an effect is due to the fact that insignificant results are rarely published.

The practice of p-hacking should have a different effect on the shape of the p-curve: authors who have successfully broken (hacked) the .05 threshold should not care much to further reduce the p-value (to, say, $p < 0.01$ or even $p < 0.001$). Thus, p-hacking should introduce a clustering of p-values just below .5 and introduce left skew into the p-curve.

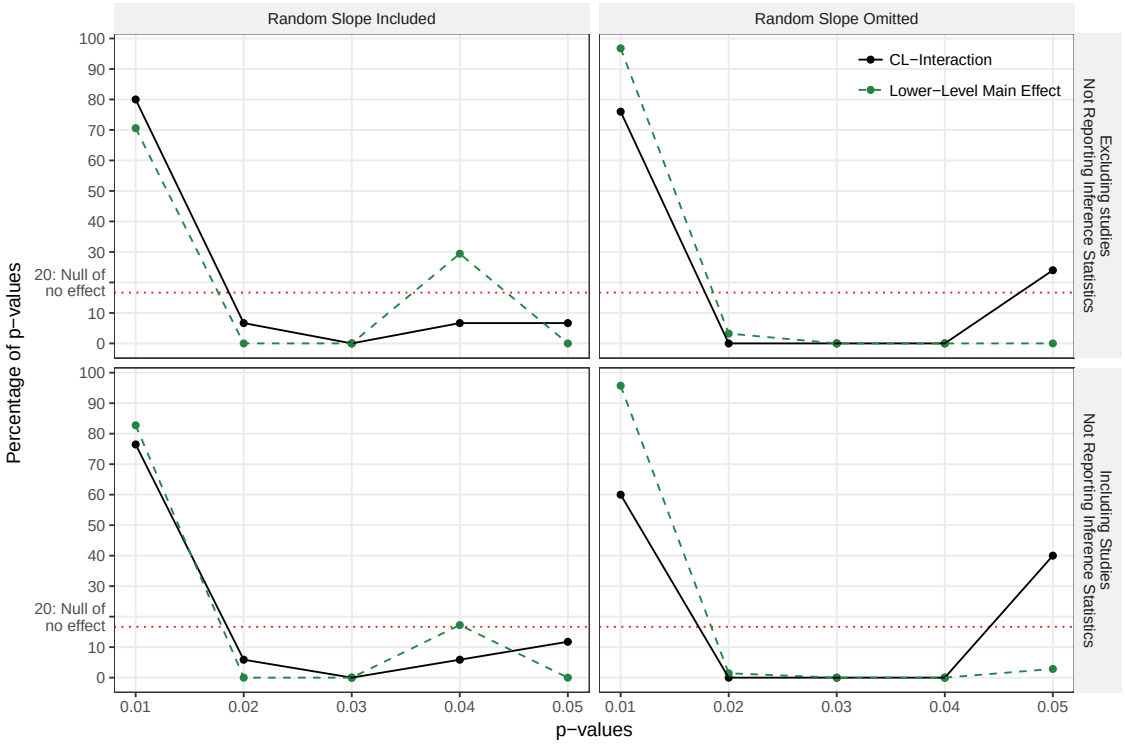
In summary, p-curves come in three principal shapes, each of which (more or less directly and convincingly) supports different conclusions concerning the evidential basis as well as the research and publication processes underlying a given collection of studies:

1. a *right-skewed* shape indicates evidential basis for a true effect;
2. a *uniform shape* indicates no evidential basis for a true effect and therefore also indicates (the potential for) publication bias;
3. a *left-skewed* shape is indicative of p-hacking and the lack of evidential basis for a true effect.

Empirical p-curves can combine these fundamental shapes. For example, a (left-skewed) p-curve with clustering of p-values below .5 and a near-uniform distribution otherwise would signal that both publication bias and p-hacking are at work. We return to this issue below.

Figure E1 displays p-curves for the cross-level interactions published in the ESR, 2011-2016. The left-hand panels show p-curves for studies that correctly include random slope terms for cross-level interactions. The right-hand panels

Figure E1: *P*-Curves for Cross-Level Interactions



Note: Results are based on 86/150 cross-level interaction terms from two-level mixed effects models for which the authors reported exact inference statistics. These were reported in 20/28 articles published in the ESR 2011-2016.

focus on studies that omitted them. The top panels show the curve for studies that allowed us to get a reasonably precise figure for the p-value, while the bottom panels also include findings for which we had to derive the p-value from an indicator, such as *. Fortunately, the shapes of the p-curves are rather robust to the in- or exclusion of studies that did not report exact inferential statistics. We will therefore focus on the top panels. The red dotted line indicates the (uniform) p-curve that we would expect to find if the results of the studies were pure artifacts of publication bias without any underlying empirical basis; it serves as the reference point for potentially right- and left-skewed p-curves. The black solid line shows the p-curve for the cross-level interaction terms and the green dashed line shows the p-curve for the main effects of the lower level variables involved in the cross-level interactions.

The four p-curves of the two top panels clearly show signs of right-skew, with the majority of p-values being smaller than 0.01. This would indicate a healthy debate based on evidential basis of truly existing associations. But the p-curves for the cross-level interaction terms also shows some indication of inflated p-values that just surpassed the threshold of the conventional level of significance ($p < 0.05$), especially for the models that omitted the random slope term in the top right panel. Simonsohn et al. (2014) suggest to test such patterns of right and left skew against the null of the uniform distribution (i.e., the red dotted line). Following their proposed method (which relies on pp-values and the Stouffer method), we learn that all four $p-$curves of the two panels are significantly right (and hence not uniform or left) skewed (all at $p < 0.0001$) and hence indicate evidential basis for real associations. If we applied the algorithm of Simonsohn et al. (2014) without further reflection, we would thus conclude that the reported findings have evidential basis and that there is no evidence of p-hacking, because all p-curves are significantly right skewed.

But in the present context such a narrow application of p-curve analysis runs into the problem that the p-curves could be both right and left skewed, that is, they could be u-shaped. This is for two reasons: first, as we do not review studies on a specific debate—but rather collections of studies that use the same modeling approach—there could be evidential basis among some and p-hacking among others, both at the same time. Second, and more importantly, a narrow interpretation of p-curve analysis has come under attack by Bruns and Ioannidis (2016), who argue that in observational studies omitted variable biases may create right skewed p-curves even in the absence of an underlying *causal* effect. We acknowledge that many of the ESR findings are not causal but associational. However, the results presented in the main article raise another serious concern. The right-hand side p-curves in Figure E1 may be right skewed simply because the omitted random slopes result in deflated p-values.

Our solution to these two problems is to exploit the following two assumptions: First, we assume that there is no systematic difference in power between studies that include and studies that omit the random slope term. Power differences might arise if one type of study investigated systematically stronger effects or worked with systematically larger samples than the other, a possibility that seems rather implausible. Second, we assume that authors potentially try to p-hack cross-level interaction terms but not the main effects of the lower-level variables. Studies that investigate cross-level interactions virtually always put the primary focus on the cross-level interaction term. The main effect of the lower-level variable, by contrast, is usually not of substantive interest. It is a conditional effect that depends on the scaling of the upper-level predictor involved in the interaction. The 'success' of an investigation of a cross-level interaction therefore primarily depends on the significance of the cross-level interaction term. At the same time, p-values for the main effects of the lower-level variable are affected by the omission of the random slope term in exactly the same way as p-values for the cross-level interaction terms. These two assumptions allow us to investigate whether the p-curves of studies that omit the random slope term are significantly more right skewed (i.e., by focusing on the lower-level main effects which are not affected by p-hacking but are similarly affected by omitting the slope term), and whether there is evidence of p-hacking (i.e., by comparing the p-curves of cross-level interaction terms against those of lower-level main effects).

Looking back at Figure E1, we can see that nearly 100% of the lower-level main effects estimated from models omitting the random slope term reach the highest levels of significance ($p < 0.01$). By contrast, among studies that correctly estimate random intercept and slope models, it is only 70%. To test whether the two p-curves are indeed significantly different from another, we employ simple dichotomous test proposed by Simonsohn et al. (2014). We transform the p-curves to a binary variable ($p < 0.025$ vs $p > 0.025$) and use a χ^2-test to investigate

whether there are statistically significant more $p < 0.025$ among studies omitting the random slope term as compared to those that include it. In principle, we could also conduct this comparison for the cross-level interaction effects. However, this comparison would be complicated by the peak of p-values near .05 for the models omitting the random slope (which is evidence of p-hacking, as discussed below). The χ^2-test comparing the p-curves for the lower-level main effects shows that the curve for models without a random slope term is significantly more right-skewed (upper panels: $p = 0.0036$; lower panels: $p = 0.0218$). This either means that these studies are better powered; as noted above, this possibility that appears quite unrealistic. An alternative—and much more likely—explanation again is that omitting the random slope term significantly deflates p-values, thus misleadingly amplifying the right skew of the p-curve. This second interpretation bolsters our claim from the main article: 'potential publication bias against insignificant findings [...] hits correctly specified cross-level interactions more often because their standard errors are not deflated'.

A final look at Figure E1 reveals another interesting comparison. In the right-hand panel (i.e., among studies that omitted the random slope term) the difference between the black solid and the green dashed p-curves (i.e., cross-level interaction terms and lower-level mains effects) shows a distinct left skew and thus indication of p-hacking. In the left-hand panel (i.e., among studies that include the random slope term), by contrast, the difference between the two p-curves seems negligible. We again use the dichotomous χ^2-test to investigate, whether this pattern is indeed statistically significant. The results are telling and unaffected by the in- or exclusion of studies that did not report exact inference statistics. Among studies that correctly specified random intercept and slope models to investigate cross-level interactions there is no significant indication of p-hacking (top panel: $p = 0.4028$; lower panel: $p = 1$). By contrast, among studies of authors who specify their models incorrectly by omitting the random slope term, we also observe

a statistically significant indication of p-hacking (top panel: $p = 0.0054$; lower panel: $p < 0.0001$). In other words: for models that omit the random slope term, there is statistically significant evidence for a higher proportion of just-significant p-values and a lower proportion of highly significant results in the cross-level interaction case than in the lower-level main effect case. We consider this as rather strong evidence for p-hacking because, as noted above, researchers usually have considerable incentive to hack the p-value for the cross-level interaction but not to hack the one for the main effect.

Appendix F Additional Monte Carlo Simulation Results

Table F1: Actual Coverage Rates (%) of Nominal 95% Confidence Interval by Number of Clusters and Lower-level Observations

		$\gamma^{(x)}$	
		Random Slope	
n_j	n_{total}	Included	Omitted
	$m = 5$ Clusters		
100	500	97.01	76.60
500	2500	96.64	43.60
1000	5000	96.59	31.81
	$m = 15$ Clusters		
100	1500	95.15	58.38
500	7500	94.89	30.52
1000	15000	95.09	21.58
	$m = 25$ Clusters		
100	2500	95.23	57.33
500	12500	94.93	29.52
1000	25000	95.01	21.03

Note: Results are based on 10,000 Monte Carlo replications. Because of Monte Carlo sampling error, the test interval is 95 ± 0.427. Values smaller or larger than that are statistically significant deviations and indicate biased inference. These results are based on experimental conditions for which $R^2(\beta_j^{(x)}) = 0.50$ (i.e., $\text{SD}(u_j^{(x)}) = 1$), and $\text{SD}(x_{ij}) = 1$.