Heisig, Jan Paul; Schaeffer, Merlin

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

European Sociological Review

Online Supplement to

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
jan.heisig@wzb.eu

Merlin Schaeffer
University of Copenhagen
mesc@soc.ku.dk

December 6, 2018

Contents

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 (Table 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 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

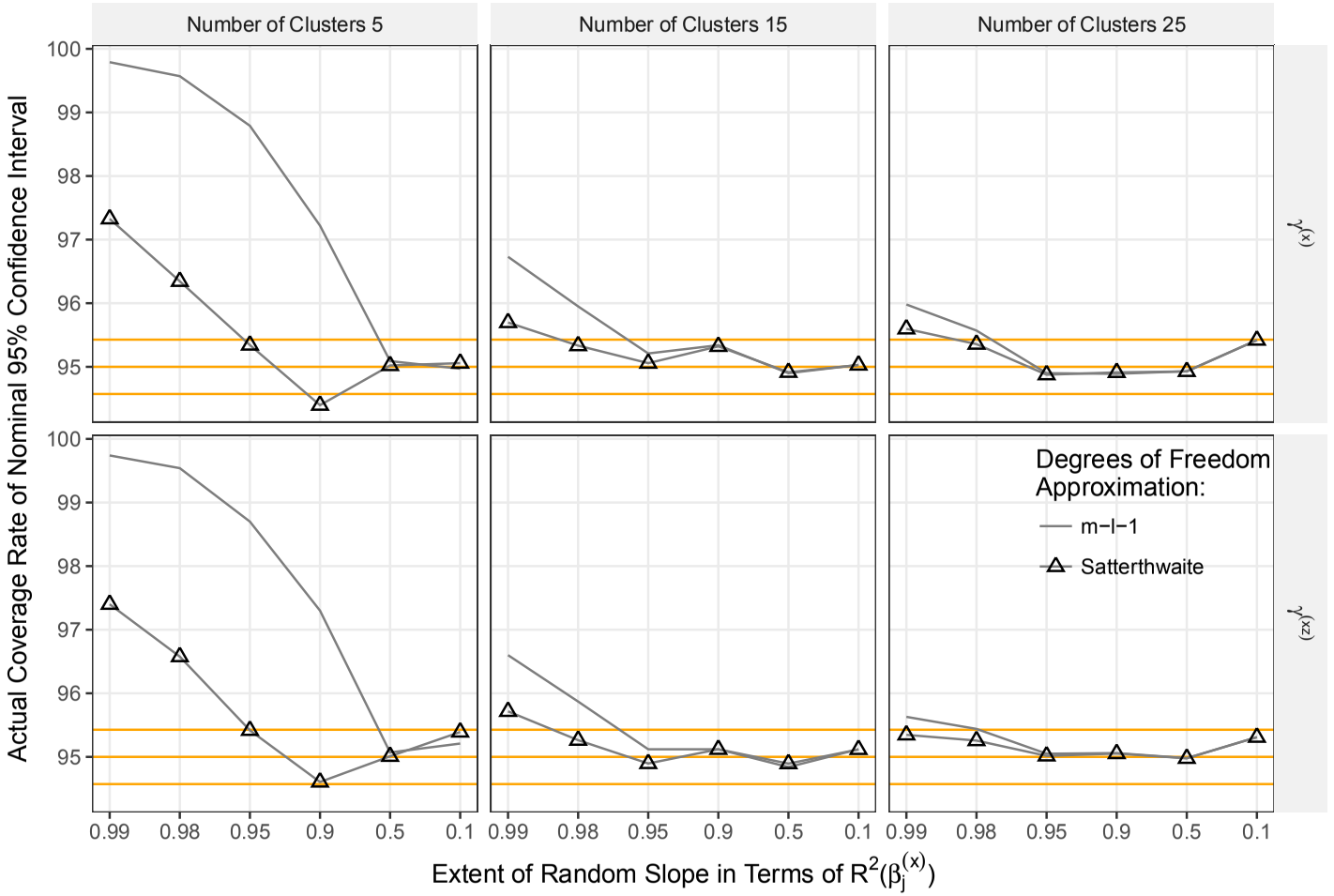
Note: Results are based on 10,000 Monte Carlo replications. Note that for reasons of brevity, this table does not express Monte Carlo error. These results are based on experimental conditions for which $R^2(\beta_j^{(x)}) = 0.50$ (i.e., SD($u_j^{(x)}$) $= 1$), and SD(x_{ij}) $= 1$.

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 over-conservative 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 over-conservative inference.[1]

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 over-conservative 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.

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.

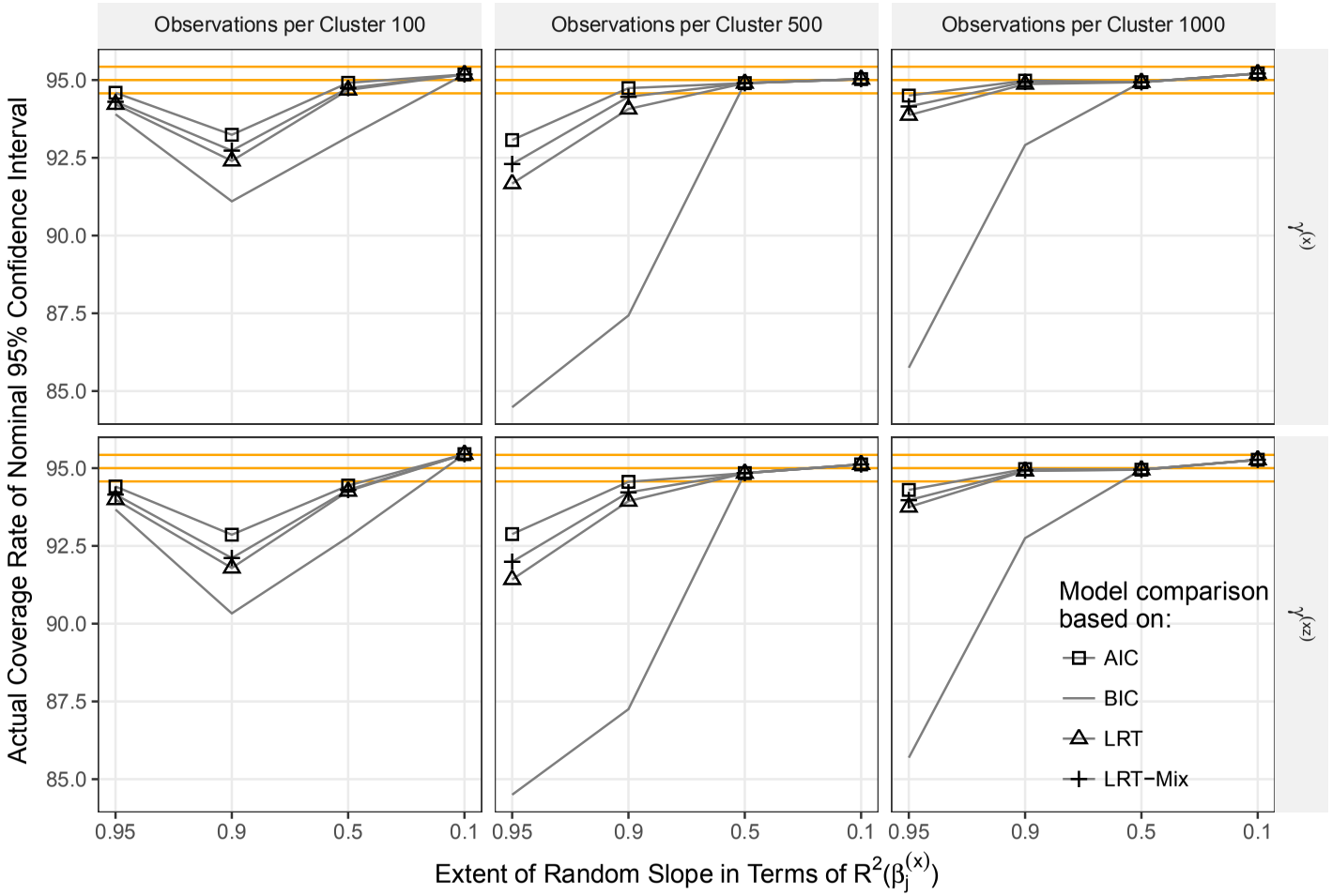
We consider four model selection criteria. The first two are variants of a likelihood ratio/deviance test (LRT). Both are based on the difference in the deviance statistic (i.e., -2 times the log likelihood) between the random intercept model and the model that includes the random slope term as well 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 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

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.

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. Intuitively, 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 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 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.

D 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 predictors 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 substantially 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. Accord-

ingly, 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 in the main article 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.

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.

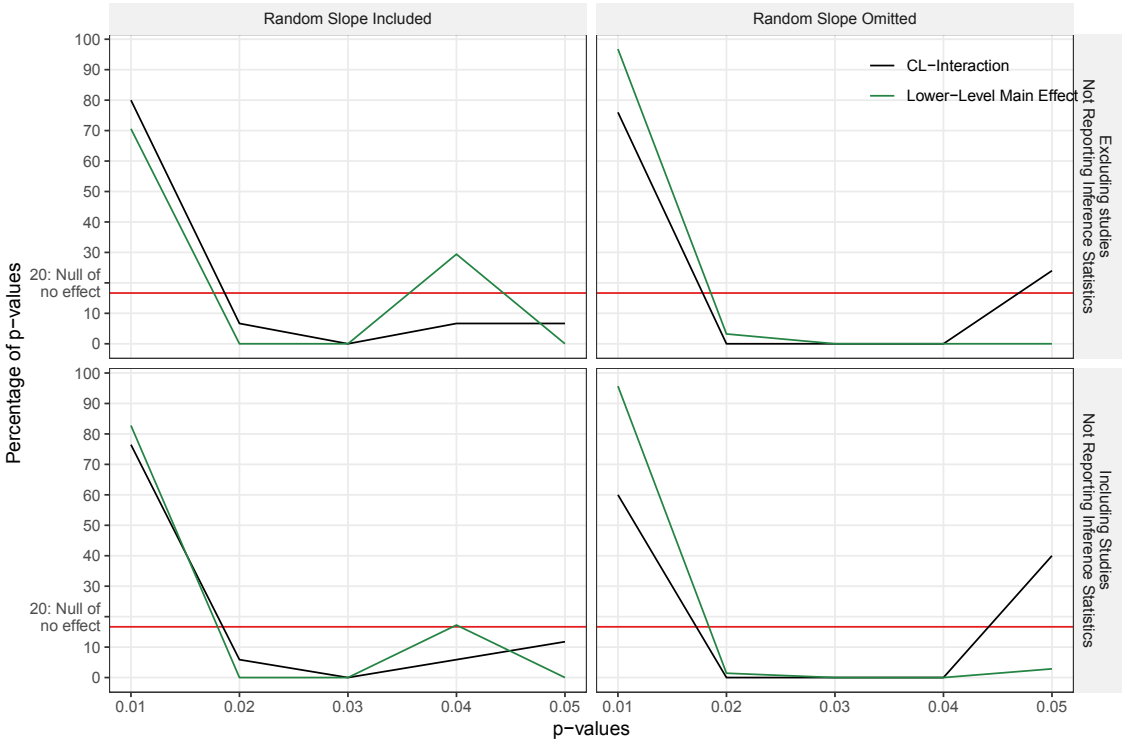
In summary, p-curves come in three principal shapes, each of which (more or less directly and convincingly) supports different conclusions concerning the evidential basis as well as the research and publication processes underlying a given collection of studies:

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.

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 95% 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 $\text{SD}(x_{ij}) = 1$.

References

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

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.

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*.

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.

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.

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**.

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.