Estimation of country-level effects in cross-national survey research using multilevel modelling: The role of statistical power

Elena Damian, Bart Meuleman, Wim van Oorschot

Introduction

Cross-national social surveys, such as Eurobarometer or World Value Survey haven been increasingly used by social scientists to explain contextual variations in people's value, opinions or behaviors for over four decades. For instance, country differences in social or political trust (e.g., Delhey and Newton, 2005; Paxton, 2007), tolerance (e.g., Hutchison and Gibler, 2007; Weldon, 2006), life satisfaction (e.g., Gundelach and Kreiner, 2004; Jagodzinski, 2010), and religiosity (e.g., Ekici and Yucel, 2014; Ruiter and van Tubergen, 2009) are just few of many issues covered using multi-national data.

An inventory of Damian, Meuleman, and van Oorschot (2019) reveals that more than half (56%) of cross-national studies published in general social sciences journals examined data by means of multilevel modeling (MLM). MLM popularity is not surprising given that it is one of the methods that enables simultaneous investigation of both individual and country-level effects. However, since the early applications of MLM, there have been constant concerns about its ability to produce accurate findings (Hox, 2010; Bryan and Jenkins, 2016). Borrowed from educational sciences – where data used is composed of students that are often nested in hundreds of classes or schools - the main concern for analyzing cross-national survey data by means of multilevel modeling is that the group-level size[1] (i.e., number of countries) might be too small to obtain unbiased and accurate effects.

As a result, there have been an increasing number of studies examining the effects of small sample sizes on the accuracy of estimates in various multilevel models used in social science research (e.g., Maas & Hox, 2004, 2005; Stegmueller, 2013; Bryan and Jenkins, 2016). A review on this issue reveals that the recommendations usually vary between 15 and 50 countries and are highly dependent on the characteristics of the models chosen, such as the type of outcome variable, whether the model has random slopes or cross-level interactions (model complexity), or the estimation procedure (see review of Bryan and Jenkins, 2016). Furthermore, a number of recent simulation studies have compared the performance of frequentist Maximum Likelihood (ML) versus Bayesian estimator procedures and show that the latter produce more accurate parameters and consequently could be a solution to the sample size issues. As a result, there is already a noticeable increase in the number of

[1] In this paper we use group-level or country-level to refer to the upper-level and individual-level for the lower-level in a two-level multilevel regression analysis.

substantive studies that are switching to Bayesian methods to the detriment of more traditional frequentist ones (McNeish, 2016; van de Schoot, 2016).

These previous research have identified important consequences of the small-N problem and proposed Bayesian estimation as a solution. Much to our surprise, however, the existing studies mainly focus on the impact of small sample sizes on parameter accuracy– i.e., relative parameter bias, relative standard error bias, and convergence rates – but rarely mention statistical power. Yet, power (that is, the probability of detecting a true population effect) is of crucial importance for accurate statistical inference and is directly linked to sample size (Cohen, 1988). Hence, the main aim of this study is to offer more comprehensive insights regarding the influence of sample size on estimation accuracy of and statistical power to detect country-level effects in multilevel regression models used in cross-national survey research. Our first research question therefore reads: To what extent does group-level sample size affect the accuracy and statistical power of country-level effects in cross-national studies? Given the rise in popularity of Bayesian methods, we also run separate models with either a frequentist Maximum Likelihood (ML) or a Bayesian estimator and analyze the differences in results. Our second research question then reads: Are Bayesian methods a solution to the consequences of small sample size found in cross-national survey research when statistical power is also considered? An additional contribution of our study is that we analyze these relations by using popular model and data structures that are commonly found in this field - i.e., two-level multilevel models with or without random slopes and cross-level interaction effects and various group-level samples). For the first time, we also examine to what extent the size of the effect of the country-level variable and the number of group-level variables in the model affects the accuracy and statistical power of the estimates. These contributions are specifically important as there are only three studies that investigate the sample size and accuracy of country-level estimates relation in the context of cross-national survey research (Stegmueller, 2013; Bryan & Jenkins, 2016; Elff, Heisig, Schaeffer and Shikamo, 2016).

Evidence on the effects of small-N problem in cross-national survey research

As multilevel regression models gained in popularity in the social sciences, various simulation studies have been examined and consequently showed that the accuracy of estimates is seriously affected by sample size (e.g. Maas and Hox, 2004; 2005; Moineddin, Matheson and Glazier, 2007; Paccagnella, 2011; Bell *et al.*, 2014). When it comes to cross-national survey research however, the conclusions from these studies are hard to generalize because of the peculiar data structure used in this field. Specifically, if in most simulations the number of individual-level observations vary usually between 5 to 65, cross-national studies use datasets with very large sample sizes per group – i.e., 500 or more.

Additionally, the models investigated in past simulations have mainly one variable at each level, while this is never the case in cross-national social research. Therefore, it is still unclear to what extent the accuracy of upper-level estimates is threatened by a low number of observations – known as the *small-N problem* (Goldthorpe, 1997). This issue is especially important here as the number of observations at the group-level is relatively low. World Values Survey has the largest sample with 60 participant countries, while most European surveys, like European Social Survey have mostly between 10 to 30 countries per survey year. In Table 1 we summarize the current (and only) simulation studies that examine, among others, the relation between group-level sample size and precision of group-level effects in the context of cross-national survey research.

Stegmueller's (2013) study aims at offering a better understanding about the necessary number of countries in typical multilevel models employed in cross-national studies (i.e., linear and nonlinear models with various levels of complexity: (1) random intercept only, (2) with added country-level predictors and/or (3) with an added cross-level interaction). His study focuses in particular on the differences in the relation between sample size and estimates accuracy across estimator procedures and consequently the findings are discussed from this angle. Stegmueller performed a simulation study with three experimental conditions: six sample sizes, three intraclass correlation values, and seven estimators (one Maximum Likelihood estimator and six different Bayesian estimators – i.e., Gibbs sampling with two different prior specifications and three posterior point summaries: expectation, median, and mode). The variation in results across different data and model conditions are evaluated by looking at the bias of estimates (percent of relative bias) and non-convergence of confidence intervals. Furthermore, the article provides detailed information about the accuracy of fixed and random parameters at each level. The findings of this study reveal that individual-level estimates are robust for both ML and Bayesian models, which is not surprising given the large number of individuals per country, usually above 1000. However, when it comes to the estimation of country-level effects, two clear differences between ML and Bayesian models can be observed. Across all sample size conditions, the relative bias of the Bayesian estimates are at most ±5%, while the ML estimates reach 10 to 15%. And with regards to confidence interval coverage, Bayesian models with country-level and cross-level interaction predictors have credible intervals that are close to the nominal level (±5%), while the confidence intervals of ML models range from 5 to 15 percentage points too narrow. The overall conclusion from this study is that ML models produce biased estimates and confidence intervals that are too short. This is particularly the case with the models with less than 15-20 countries and/or models with a higher number of variables. By comparison, Bayesian models produce robust findings and are therefore recommended when sample size is a concern. Yet, Stegmueller (2013) points out to readers that in the case of increasing the

number of macro or cross-level interaction variables – i.e., increase in model complexity - Bayesian models will also be prone to produce unreliable results.

In a follow up study, Elff, Heisig, Schaeffer and Shikamo (2016) compare the performance of different estimation procedures in relation with the sample size issue as well. They employ the exact model and data experimental conditions as Stegmueller, but add to the mix a new condition using a frequentist REML estimator and construct the confidence intervals using the t-distribution with degrees of freedom based on the number of groups[2]. Their simulations reveal that even with a sample size of five countries, frequentist REML models will produce accurate estimates and illustrate that it is possible to obtain accurate estimates without turning to Bayesian methods.

Overall, the findings from these two studies confirm that the small the number of countries found in cross-national survey research can seriously affect the accuracy of group-level estimates, but this issue seems to be resolved by means of refining estimation procedures. However, these as well as most similar other simulations have been focusing solely on the accuracy of parameter estimates, measured by relative bias and coverage rates. What is still missing is information about the relation between sample size and statistical power. Looking at statistical power is crucial as it represents the probability of correctly concluding significance when there is a true effect in the population. It is a function of the level of significance (alpha level) - usually set to .05, the strength of association between two variables – size of the effect -, and the sample size at each level (Cohen, 1988). Therefore, including statistical power as one of the evaluation criteria is of paramount importance to understand the consequences of using small sample sizes on the robustness of the country-level effects. In other words, obtaining unbiased parameters is not enough if there is not sufficient statistical power to estimate them. Furthermore, it remains unknown whether Bayesian methods can still be a solution to the consequences of the *small N problem*, and if that is the case, in what circumstances/scenarios/for what types of model and data combinations.

Another issue with previous studies is that the model conditions investigated differ greatly from the models employed in cross-national research. The study of Bryan & Jenkins' (2016) tackle this issue and provide some useful insights into the relation between sample size, model complexity, and accuracy of estimates, by using data and model conditions closer to what we find in "real life cross-national studies". They examine accuracy of estimates in both linear ML and logistic REML models that have several variables per level, a mix of binary, categorical, and continuous regressors, and include some extended models with two random slopes and cross-level interactions added as well. Contrary

[2] "At least three textbooks on multilevel mixed effects models mention them *m - l - 1* approximation to the degrees of freedom for testing context effects (Pinheiro and Bates 2000, 91-92; Raudenbush and Bryk 2002, 57-58; Snijders and Bosker 2012, 94-95). As before, *m* denotes the number of clusters, l the number of contextual effects, and 1 is added for the intercept."(Elff et al., 2016, p.17)

to previous studies, the values of the parameter estimates are chosen by first running a set of multilevel models on European Union Statistics on Income and Living Conditions survey data - the dependent variables are either hours of work or labor market probabilities and independent variables are gender, education, or GDP per capita (see full list in the original paper). As accuracy criteria, they use relative parameter bias, the 95 per cent CIs for relative bias statistics (and Root Mean Squared Error), relative SE error, and non-convergence rate. The results indicate that a minimum sample size of 25 or 30 countries is necessary to obtain robust estimates in a linear and logistic MLM respectively. Furthermore, another interesting finding is that adding cross-level interactions and random slopes increases the complexity of the models and consequently the bias of the country-level estimates. Therefore, similar to Stegmueller (2013), Bryan & Jenkins (2016) warn readers that a higher sample size might be needed if readers are to use more complex models - i.e., higher number of parameters - than the ones used in their simulation study.

Table 1. Monte Carlo simulations studies about the influence of small sample size on accuracy of country-level effects in cross-national research

Study	**Stegmueller (2013)**	**Elff et al. (2016)**	**Bryan & Jenkins (2016)**
Relation of interest	Relation between sample size and accuracy of country-level estimates	Critique and extension of Stegmueller's (2013) study	Relation between sample size and accuracy of country-level estimates
MLM structure	1 level-1, 1 level-2, with or without cross-level interaction effect	1 level-1, 1 level-2, with or without cross-level interaction	7 level-1, 1 level-2
Effect size of level-2 variable(s)	.43	0.43	-.23*
n_j	500	500	1000
N	5, 10, 15, 20, 25, 30	5, 10, 15, 20, 25, 30	5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100
Estimator method	ML and Bayesian	ML, REML, and Bayesian	ML (linear models) and REML (for logit models)
Assessment criteria of level-2 estimates	Relative parameter bias, coverage of the 95% confidence interval	Relative parameter bias, coverage of the 95% confidence interval	Relative parameter bias, standard error bias, coverage of the 95% confidence interval

Studying what happens when more variables at the upper-level are present in the model – i.e., the complexity of the model increases - is particularly relevant because of the constant confusion

about the ratio between the sample size and the number of variables allowed in a multilevel model. For instance, we often see papers where scholars argue that by running and presenting the findings of models with one country-level variable at a time, they account for the possible negative effects of the small sample size. To our knowledge, this relation has not been tested yet. Furthermore, most simulation studies look at the estimation accuracy of a fixed effect size, usually medium to large, while the accuracy and statistical power of small to medium range has not been examined (.10 and .30, according to Cohen's classification, 1982). We therefore believe that in order to get a deeper understanding regarding the consequences of sample size on accuracy of multilevel estimates, it is necessary to examine what happens when the size of the effect varies as well.

Specifications of the simulation study

With this study we analyze whether the number of participant countries found in cross-national survey research is large enough to detect a true country-level effect (i.e., if there is enough statistical power to find such effects). The second general aim is to examine the possible variation in the relation between sample size and parameter robustness across two well-known estimation procedures: frequentist ML and Bayesian. Specifically, we investigate whether Bayesian methods are still a solution to the small sample size when statistical power is taken into account. We answer these questions by performing a Monte Carlo simulation procedure (Muthen & Muthen, 2002) with model and data conditions commonly found in cross-national survey research.

1. Baseline: A Random intercept model with individual and group-level predictors

Our baseline model is a two-level linear hierarchical model with Y_{ij} representing a continuous outcome variable (the subscript *i* is for individuals and *j* for the groups), γ_{00} is the intercept, X_{ij} - X_{ij} individual-level explanatory variables, Z_{1j} a group-level explanatory variable with a medium standardized effect (i.e., .25), u_{0j} and e_{ij} are the residuals at the lower and upper level respectively[3] (Equation 1).

$$Y_{ij} = \gamma_{00} + \sum_{P=1-5} \gamma_{o_P} X_{0p} + \gamma_{01} Z_{1j} + u_{0j} + e_{ij};\ i = 1 \ldots n_j,\ j = 1 \ldots J \quad (1)$$

The country-level variable Z_1 was our variable of interest and we tracked the changes of various accuracy indicators across the following model specifications:

[3] Given the different systems of notations used in the multilevel literature, we want to specify that in this paper we use the ones of Hox (2010). Furthermore, we do not provide introductory explanations of what a multilevel analysis is as this is not the goal of the paper. However, for more information about this type of analysis, readers are encouraged to consult Hox (2010) or Raudenbush & Bryk (2002).

- country level effect size: .10, .25, and .50
- country-level sample size: 15, 20, 30, and 40
- estimator: Maximum Likelihood (ML) and Bayes
- model complexity: with one to five country-level effects

Unlike most simulations that investigate the accuracy of group-level variables with medium to large effects (see for example Table 1), we looked at differences in accuracy of estimating low, medium, and large country-level effects (classification of Cohen, 1988, 1992) across different model conditions. The sample size conditions were chosen based on the size of the cross-national social surveys that usually contain 20 to 30 countries. We included models with only 15 countries as well because, sometimes, due to missing values or focus of the study, researchers have to restrict their study sample size to 20 or even fewer countries. The sample size at the individual level remains fixed to 1,000 as survey data contain large number of individuals per country and there are no major accuracy issues. For the third experimental factor, we increased the baseline model complexity by adding up to four additional country-level variables. And, given the literature that discusses the benefits of using a Bayesian multilevel models, (Stegmueller, 2013), we run separate models with a Full Maximum Likelihood and a Bayesian estimation procedure. Concerning the former, we used the default prior distributions of Mplus because this is the common practice in substantive research (Smid, McNeish, Miočević & van de Schoot, 2019) and specified a convergence criterion equal to .1.

All parameter values were chosen to reflect what is found in typical cross-national social studies. The individual-level explanatory variables (X_1 – X_5) and the additional four country-level variables (Z_2–Z_5) took a medium standardized effect of .25 (Cohen, 1988). All models with two or more group-level variables included a medium correlation between them as well and for all models we used an intra-class correlation equal to .10 – this is considered to be a medium size for the cross-national research (Stegmueller, 2013).

2. Two additional models: Random slope model and cross-level interaction model

We run an additional set of multilevel models with a random slope and cross-level interaction as these are also employed to analyze cross-national survey data. To the baseline model, we added a random effect of level-1 variable X_1 and, in a follow up step, either a medium or high cross-level interaction effect between X_1 and Z_1 (i.e., .25, .50). Here, we had three experimental conditions: sample size, estimation procedure, and size of cross-level interaction effect. All models with a random intercept and slope included a medium size correlation as well.

In total, we had 144,000 simulated conditions. One hundred twenty experimental conditions for the first model (4 sample sizes X 3 effect sizes X 5 combinations of variables X 2 estimators), 24 experimental conditions for the two additional models (4 sample sizes X 3 cross-level interaction effect sizes X 2 estimators), and 1,000 simulations in each of these conditions. Furthermore, the estimation algorithm converged for all estimated models. Appendix 1a and 1b provides an overview of the model specifications.

Estimation criteria of country-level effects

We used three criteria to assess the accuracy of the parameter estimates - relative parameter bias, relative standard error bias, coverage rates - and statistical power. T*he relative parameter bias* is calculated by subtracting the population parameter value from the average parameter estimate over the 1,000 replications of the simulation study and then divide it by the population value. *Relative standard error bias* is calculated in a similar way, with the exception that we used the standard deviation of each parameter estimate over the replications as a measure of the population standard error. This is considered to be a reliable approximation of the population standard error when the number of replications is large (Muthen and Muthen, 2002). According to Hoogland and Boomsma (1998), for an accurate parameter estimation, a bias value of +/- .5 (5%) is acceptable, while the maximum value of relative standard error bias should not exceed +/- .10 (10%). The c*overage rate* represents the proportion of replications for which the 95% confidence or credible interval holds the population parameter. The *empirical power* of statistical tests of explanatory variables represents the percentage of replications for which we can reject the null hypothesis that there is no effect at the conventional alpha value of .05 level (two-tailed test, Z_{crit}=1.96; Muthen & Muthen, 2002, p. 606). According to Cohen (1992) the likelihood of detecting a significant population effect different from zero needs to be 80% or higher.

Results

The focus of this study and consequently the conversation in this section is on the estimation accuracy and statistical power of country-level effects. To keep the results section coherent and easy to follow, we focus on discussing and displaying the findings of our baseline model – random intercept model with one country-level variable (Table 2) – as well as some of the additional models with a random slope and a medium cross-level interaction effect (Table 3). Furthermore, the statistical power findings of the random intercept model across all experimental conditions, including the model complexity condition, are displayed in Figure 1. The full results tables of our simulation study are provided in Appendix 2 and Appendix 3.

1. Random intercept only models

We start by discussing the influence of sample size, effect size and estimation procedure on accuracy of country-level effects in frequentist Maximum Likelihood models. In Table 2 we can see that the model with 15 countries and a low effect size (.10) registers a downward parameter bias of 10%. Given the acceptable bias of 5% (Hoogland and Boomsma, 1998), the parameter estimates for this type of model are inaccurate. Similarly, the standard error bias is about 14% too small and as a consequence the coverage rate is estimated too short as well (.88). For only 88% of replications the 95% confidence interval coverage rate contains the population value. As a reference, a standard error bias below 10% and a coverage rate around 95% are desirable for accurate estimates. Furthermore, one minus the confidence interval coverage rate is the nominal alpha value of .05 or the probably of wrongly concluding that there is a population effect (Type I error). This means that when the coverage rates are too short, the alpha value is inflated.

Concerning the influence of sample size, our findings confirm that the accuracy of our estimates improves with sample size. Specifically, in the models with a low effect size but with 20, 30 and 40 countries, the parameter estimate is underestimated by 11, 2, and 1% respectively. Hence, the bias becomes negligible once the group-level sample size reaches 30. Similarly, the S.E. error bias is well within the acceptable limits (-.07) in the models with higher sample sizes (≥30). The coverage improves also but only reaches 92% in the model with 40 countries.

As excepted, we also find an association between the size of the effect and accuracy of estimates. Starting again with the ML models with 15 countries, we can see that the parameter bias for the low (.10), medium (.25) and high effect (.50) is -.10, -.04 and -.02 respectively. And an almost identical downward trend is observed within the other sample size conditions. Looking at the influence of both sample size and effect size, we find that the estimates are seriously biased only when both the effect and sample size are on the small side. The size of the effect does not have an influence on the S.E. bias and coverage rates of the country estimate.

Going further to the differences between ML and Bayesian models, our simulations show that the latter produce more accurate estimates, which is in line with previous research (Stegmueller, 2013; Bryan & Jenkins, 2016). If we take the model with a small sample size (15) and low effect size (.10), we can see that the parameter bias is -.07, standard error bias is .17, and coverage rate is .96 in the Bayesian model, while the corresponding values in the ML model are -.10, -.07, and .88. The parameter bias is slightly lower than in the ML models and once the sample size and/or effect size increase, the bias drops below the 5% acceptable bias. The standard error is overestimated, but again the bias decreases to acceptable values once the number of countries increases to 20. Furthermore,

the coverage rates are much closer to the nominal rate compared to the ML models – i.e. between 94 and 96 % compared to 88 to 92%.

The findings regarding the statistical power to estimate a country-level effect tells a different story. Looking at the ML models with 15 countries, the statistical power to detect a small effect size is only .13 and does not register a remarkable increase when larger sample sizes are used. Specifically, in the model with 40 countries, the statistical power is merely .12. Going further to the models with higher effect sizes, the statistical power for a medium size effect is between .23 to .38 across the country-level condition and between .60 to .93 when a large effect size is specified. This means that even when using a large cross-national survey data (a sample size of 40 countries is relatively large in this field), the probability of detecting a true population small and medium effect is only 12% and 23% respectively. In other words, there is a high probability of false negatives (i.e., Type II error) and consequently wrongly concluding that there is an effect. Another surprising result here is that we find no large differences between the two estimation procedures when it comes to statistical power. The Bayesian models have actually lower statistical power to detect a contextual effect compared to the ML ones. The statistical power to detect a small effect in the ML models with 15 countries is .13 and only .05 in the Bayesian models. Furthermore, similar to ML models, only when both the sample size (>30) and effect size (≥.50) are large, the probability of detecting a true population effect is close to the .80 threshold. It is important to mention here that even though the ML models have higher statistical power to detect country-level effects, this is because these models do not fare well with small sample sizes (Browne & Draper, 2006; Lee & Song, 2004; McNeish & Stapleton, 2014, McNeish, 2016). For instance, if we look at the standard error bias for the same effect, we have 10% and 7% bias in ML and Bayesian models respectively. Thus, the statistical power is higher but the chance of a false positive is high as well. To further demonstrate this, we also run a few models with either a population effect zero or cross-level interaction effect zero (Appendix 4). These results confirm that Maximum Likelihood models have a higher probability of finding significance when there is no effect in the population. For instance, in a random intercept model with 15 countries, the probability of wrongly concluding significance is 12%, while in a Bayesian model this is only 3%. As the sample size increases, the difference in power estimates diminishes. This is explained by the fact that the standard errors tend to be underestimated in ML models with small sample sizes (see the standard errors in Table 2). Altogether, our results illustrate that both the accuracy and statistical power of country-level effects are affected by sample size, effect size, and estimation procedure. However, if the issue of accuracy seems to be solved by using sample sizes larger than 20, this cannot be said for the statistical power issue. Additionally, these results confirm that the size of the effect needs to be taken into account when looking at the relation between power and sample size.

Table 2. Random intercept model with one country-level variable: Estimation criteria for various effect sizes, country-level sample sizes, and estimation procedures – 24,000 replications

		Low effect size				**Medium effect size**				**Large effect size**			
		15	20	30	40	15	20	30	40	15	20	30	40
Parameter bias	Maximum Likelihood	-.10	-.11	-.02	-.01	-.04	-.04	-.01	.00	-.02	-.02	.00	.00
	Bayesian	-.07	-.02	.03	-.01	-.03	-.01	.01	-.01	-.01	.00	.01	.00
S.E. bias	Maximum Likelihood	-.14	-.11	-.07	-.07	-.14	-.11	-.07	-.07	-.14	-.11	-.07	-.07
	Bayesian	.17	.10	.05	.01	.17	.10	.06	.01	.17	.11	.06	.01
Coverage rates	Maximum Likelihood	.88	.91	.91	.92	.88	.91	.91	.92	.88	.91	.91	.92
	Bayesian	.96	.96	.95	.94	.96	.95	.95	.94	.97	.95	.95	.95
Statistical power	Maximum Likelihood	.13	.11	.11	.12	.23	.24	.31	.38	.60	.70	.86	.93
	Bayesian	.05	.05	.09	.10	.12	.15	.26	.32	.42	.58	.77	.90

As a last step in our study, we were also interested to see to what extent accuracy and statistical power of country-level estimates are affected by the complexity of the model. We studied this relation in two scenarios/settings. Firstly, we increased the complexity of our random intercept models by adding up to four additional country-level variables and, secondly, by using the same baseline model - with one medium effect size variable -, we run some additional models with a random slope and cross-level interaction variables. The results for the latter are discussed in the following subsection.

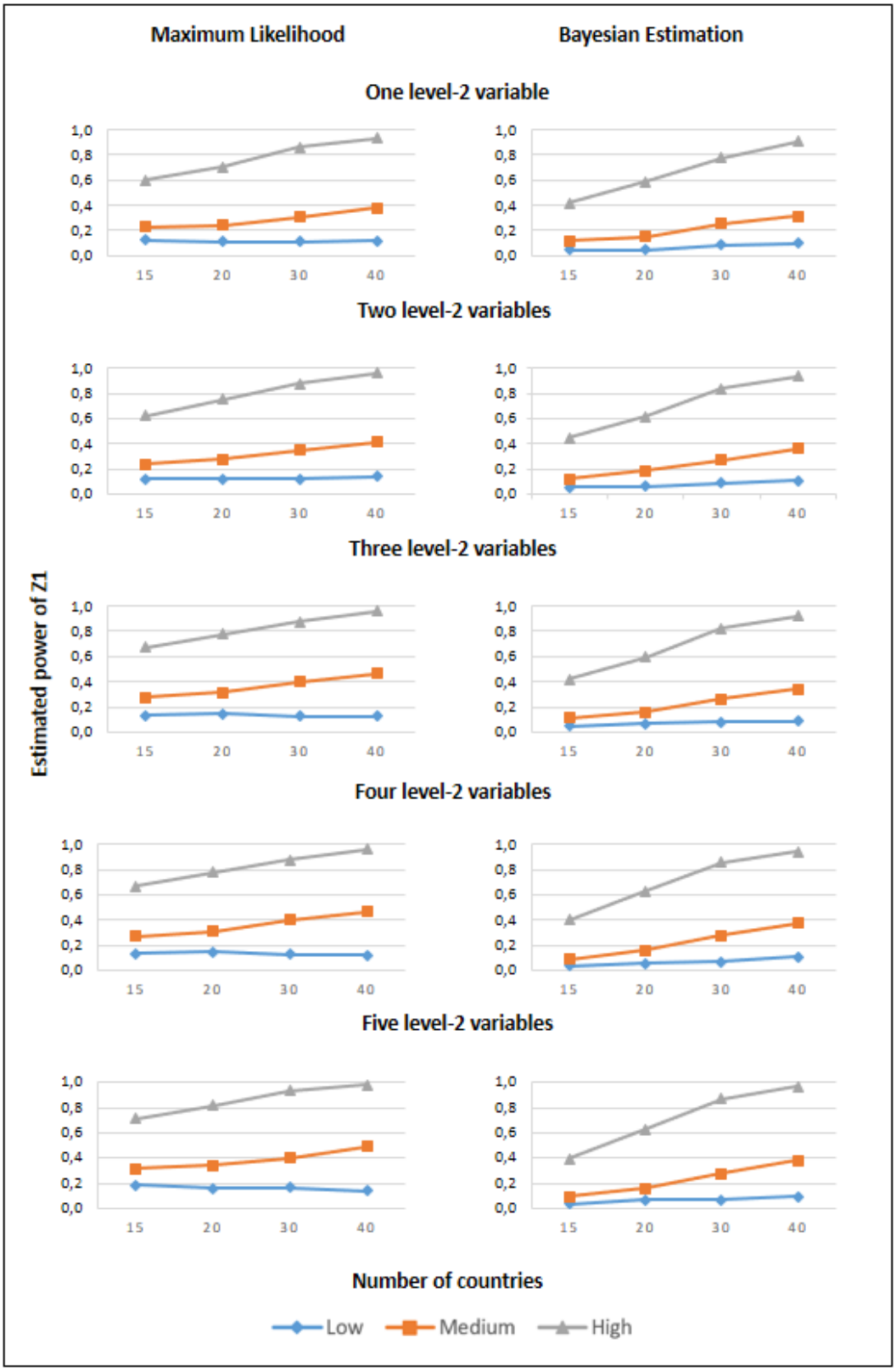
The findings for the first set of models can be found in Appendix 2. Regarding parameter bias, we can see that the ML models with one country-variable and a low effect size have some bias when the sample size is either 15 or 20 (i.e. -.10 or -.11), but this bias rapidly diminishes when additional same-level variables are added to the model. Specifically, the model with 15 countries, a low effect size, and no other variables has a the parameter bias of -.10, which drops to -.02 in the model with one additional variable, to .03 in the model with either two or three extra variables, and to -.03 in the model with four additional variables. If we look at the models with higher effect sizes, we can see a similar decrease in bias when more variables are added to the model. Furthermore, compared to the low effect size models, here the parameter bias is within the 5% limits across all models with different number of country-level variables. Regarding the models with a Bayesian estimator, the parameter bias is much lower and stays relatively stable across the model complexity condition, compared to the ML models.

The standard error bias for the ML models with a low effect size and a sample size between 15 and 30 increases as we add more variables to the model (Appendix 2b). The same pattern can be observed in the models with medium or high effects. If we are to compare the ML model with one country-level variable and a low effect size to the same one but with four additional variables (Model 5), we can see that the standard error bias doubles. Thus, an increase in complexity of the Maximum Likelihood models does affect the accuracy of its country-level estimates. Specifically, the standard errors are seriously underestimated which leads to unreliable findings. Furthermore, the relation between accuracy and model complexity stabilizes when the number of countries increases. We can see that the standard error bias for the models with 40 countries stays quite stable across the effect size and model complexity condition and has acceptable accuracy – between -0.8 and -.05. Regarding differences between estimation procedures, in the Bayesian models, the standard errors bias tends to increase with the number of added variables as well, however this happens only when the sample size is on the small side (15 or 20 countries). Besides these models, all other Bayesian model conditions have no notable bias.

We find minor to no differences in terms of coverage rates across the model complexity and effect size conditions. There is a slight decrease in coverage rate as the number of variable added increases, and this can be explained by the poorly estimated standard errors. Specifically, if we look at the ML model with 15 countries and a low effect size, we can see that the coverage rates range between 88% in the one-variable model to 94% in the five-variable model. The coverage rate stays the same regardless of the size of the effect. Furthermore, for the same model, we ca notice an improvement in coverage with samples size, namely from .88 to .92 in the model with 40 countries. Regarding the influence of estimator procedure, Bayesian models have better coverage compared to the ML models. For example, in the model with 15 countries and one country-level variable with a low effect, the coverage rates for the Bayesian and ML models are .96 and .88. Furthermore, the coverage rates are more stable across all model complexity conditions.

In Figure 1 we illustrate the variation in statistical power to detect a country-level effect across all experimental conditions. Overall, statistical power increases with sample size, effect size, and model complexity in both ML and Bayesian models. Regarding model complexity, it is striking to see that the estimated statistical power does not suffer notable changes when additional variables are added to the model. In the ML model with 15 countries, one country-level variable, and low effect size, the statistical power to detect a low effect is .13 and increases to .20 in the model with five same-level variables (Figure 1, first and last left-side graphs, Appendix 2d for the exact values). A similar pattern can be observed for the same models with a medium or large effect. Thus, the models with fewer variables do not have higher power to detect a certain effect and therefore providing additional models with one variable at a time is not a strategy to control for a small sample size. On the contrary, if we look carefully at the power estimates of the models with one and five group-level variables (Appendix 2d provides the exact values), we can notice that the estimates of the later are higher. This means that adding more variables to the model can actually improve the precision of the variable of interest. By all means, additional variables should only be added when there is theoretical reasoning for it and if they explain some of the variation between countries or groups.

Figure 1. Power for detecting a country-level effect (Z_1) – various country-level sample sizes, effect sizes, estimation procedures, and number of level-2 variables – 120,000 replications



2. Random slope model and cross-level interaction model

In a second series of simulations, we evaluated the estimation accuracy and statistical power of a medium-size main country-level effect in models with a random slope and cross-level interactions (see Appendix 3 for full results). Table 3 shows the findings from the models with both a medium–size main and a cross-level interaction effect. Looking at the findings for the main country-level effect, the ML models with small samples sizes (15 or 20) tend to produce estimates with higher bias and narrower coverage rates. Specifically, in the 15 countries condition, there is an initial parameter bias of .06, which completely disappears when the sample size is equal to 40. Furthermore, a similar descending trend can be noticed for the S.E. bias, while the coverage rate increases from .91 to .94. The estimated power to find a medium-size effect lies between .24 and .38 and, therefore similar to the findings from the random intercept only models, the probability of detecting a true population effect lies far below 80%. Specifically, even when using a large cross-national survey dataset (i.e., 40 countries), the statistical power is only .38. What it is worth mentioning here is that power to estimate a country-level effect stays the same regardless if the cross-level interaction effect is large and consequently has a high power. These results are shown in Appendix 3d.

Table 3. Random slope and cross-level interaction model: Estimation criteria of country-level estimates for various country-level sample sizes, estimation procedures, and model complexity – 8,000 replications

		Medium effect size				**Medium cross-level interaction effect size**			
		15	20	30	40	15	20	30	40
Parameter bias	Maximum likelihood	.06	.05	.02	.00	.00	.04	.02	.02
	Bayesian	-.03	.01	.00	.03	-.01	-.01	-.01	.01
S.E. bias	Maximum likelihood	-.07	-.06	-.05	-.04	-.12	-.06	-.05	-.03
	Bayesian	.36	.19	.07	.04	.36	.25	.14	.07
Coverage rates	Maximum likelihood	.91	.92	.92	.94	.89	.94	.93	.94
	Bayesian	.98	.98	.97	.96	.98	.98	.96	.96
Statistical power	Maximum likelihood	.24	.27	.31	.38	.20	.23	.28	.34
	Bayesian	.07	.12	.24	.35	.06	.10	.18	.26

Regarding the differences in accuracy between the two estimation procedures, the estimates obtained using a Bayesian approach are more accurate parameter estimates compared to the ML ones. Namely, the parameter bias fluctuate between -.03 and .01 across the samples size condition. These are however negligible differences as both estimation procedures register values close to or

around the conventionally accepted limits of ±.05. Furthermore, the standard error tends to be underestimated in the ML models and overestimated in the Bayesian ones. In both situations, the accuracy increases with samples size. For instance, in the Bayesian model, the S.E bias decreases from .36 in the model with 15 countries to .04 in the model with 40 countries. Concerning statistical power, we find no large differences between the types of models. For example, even when using a large cross-national survey dataset (i.e., 40 countries), the statistical power is only 38% and 35% in the ML and Bayesian model respectively.

Discussion and conclusions

Even though multilevel regression analysis has been frequently used in social science research for the last decades, there is still confusion regarding its ability to produce robust estimates when relatively small sample sizes are available. An increasing number of simulation-based research has been providing useful insights on this topic (see review of Bryan and Jenkins, 2016; Smid *et al.*, 2019), however, because the model and data structures found in cross-national survey are too distinct from what is usually found in the social sciences (i.e. small groups and very large number of observations within group), the recommendations offered are difficult to generalize. Therefore, the aim of this paper was to analyze whether the group-level sample sizes, usually number of countries can seriously threaten the robustness of group-level effects and consequently whether these multi-national survey data can be safely examined by means of multilevel modeling. One of the main contributions of our studies is that, unlike previous studies, we looked at both accuracy and statistical power of group-level estimates found in commonly used multilevel models. Furthermore, besides the influence of sample size, we also investigated to what extent the estimation procedure, effect size, and model complexity play a role in the precision of country-level estimates.

Firstly, a striking finding of our simulations is that the group-level sample sizes found in cross-level survey research are too small to detect a small or medium size group-level effect. Particularly, we find that only when the group-level sample size is equal to 30 or higher and the country level variable has a high effect (≥.50), the statistical power to detect such an effect reaches the conventional level of 80%. Additionally, this result highlights that simulation studies that analyze the effects of sample size on estimation of country level effects should also include an experimental condition with the effect size and use statistical power as one of the main estimation criteria.

Secondly, in line with previous research, we find that the use of Bayesian methods results in more accurate estimates in terms of parameter bias, standard error bias, and coverage rates compared to the traditional ML models, yet the differences between the two estimation procedures are not outstanding. The results indicate that the models with both small sample sizes – i.e., 15 or 20

observations in our study - and small effect size, have the lowest accuracy of parameters and, as expected, the ML models perform worse than the Bayesian ones. One explanation for these differences is that ML estimation procedure fares worse with very small sample sizes. Still, even if Bayesian methods produced unbiased estimates and standard errors with a lower number of group-level observations, this does not guarantee that our models will have enough power to estimate these effects. Therefore, the combination of the four assessment criteria and especially the inclusion of power provides a much clearer picture of how group-level sample sizes influence the accuracy of country-level estimates and their standard errors.

Finally, for the first time we show that the complexity of the model does not have a strong impact on the robustness of the country level parameters. This finding contradicts current beliefs that running models with one variable at a time might diminish the negative effects of a small sample size. Furthermore, upon closer investigation, we found that the precision of estimating an effect, increases when more variables are present. Specifically, when models contain a higher number of variables and that explain some of the between-country variance, the precision of estimating a country-level variable increases. We also run some additional models with random slopes and cross-level interactions and found again that what really has an effect on the accuracy of estimates is the sample size and size of effects, and not the model complexity.

Overall, these results reveal that applying MLM to cross-national data is likely to result in under-powered studies because of the low sample sizes and/or the prevalence of small effects in this field. Furthermore, it highlights the importance of using Monte Carlo simulation study for post-hoc power analyses and report whether the models performed have enough power to estimate the country-level effects found.

References

Bell, B. A. et al. (2014). How low can you go? An investigation of the influence of sample size and model complexity on point and interval estimates in two-level linear models. Methodology 10, 1–11.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood based methods for fitting multilevel models. Bayesian Analysis, 1, 473–514. doi:10.1214/06-BA117

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

Cohen, J. (1988). Statistical power analysis for the behavioral sciences, 2nd ed. New York: Routledge.

Cohen, J. (1992). Quantitative methods in psychology: a power primer. Psychological Bulletin, 112(11), 155-159.

Delhey, J. & Newton, K. Predicting cross-national levels of social trust: Global pattern or Nordic exceptionalism? European Sociological Review, 21(4): 311-327.

Gundelach, P. and Kreiner, S. (2004). Happiness and life satisfaction in advanced European countries. Cross-Cultural Research, 38 (4): 359-386.

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.

Ekici, T. and Yucel, D. (2014). What determines religious and racial prejudice in Europe? The effects of religiosity and trust. Social Indicators Research, 122(1): 105-133.

Goldthorpe, J.H. (1997). Current issues in comparative macrosociology: a debate on methodological issues. Comparative social research, 16, 1-26.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. An overview and a meta-analysis. Sociological Methods & Research, 26(3), 329-367.

Hox, J. J. (2012). Multilevel analysis: Techniques and applications, 2nd ed. London: Routledge.

Hutchison, M.L. and Gibler, D.M. (2007). Political tolerance and territorial threat: A cross-national study. Journal of Politics, 69(1): 128-142.

Jagodzinski, W. (2010). Economic, social, and cultural determinants of life satisfaction: Are there differences between Asia and Europe? Social Indicators Research, 97 (1): 85-104.

Lee, S. Y., & Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. Multivariate Behavioral Research, 39, 653–686. doi:10.1207/s15327906mbr3904_4

Maas, C. J. M. and Hox, J. J. (2004). Robustness issues in multilevel regression analysis. Statistica Neerlandica, 58, 127–137.

Maas, C. J. M. and Hox, J. J. (2005). Sufficient sample sizes for multilevel modelling. Methodology, 1, 86–92.

McNeish, D., & Stapleton, L. M. (2014). The effect of small sample size on two level model estimates: A review and illustration. Educational Psychology Review, 28, 295–314. doi:10.1007/s10648-014-9287-x

Mc Neish, D. (2016). On Using Bayesian Methods to Address Small Sample Problems, Structural Equation Modeling: A Multidisciplinary Journal, 23:5, 750-773, DOI: 10.1080/10705511.2016.1186549

Moineddin, R., Matheson, F. I. and Glazier, R. H. (2007). A simulation study of sample size in multilevel regression models. BMC Medical Research Methodology, 7, article 34.

Muthen, L. K., & Muthen, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. Structural Equation Modeling, 9(4), 599-620.

Paccagnella, O. (2011). Sample size and accuracy of estimates in multilevel models. Methodology, 7, 111–120.

Paxton, P. (2007). Association memberships and generalized trust: A multilevel model across 31 countries. Social Forces, 86(1): 47-76.

Raudenbush, S. W. and Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods, 2nd ed. Thousand Oaks, CA: Sage Publications.

Ruiter, S. and van Tubergen, F. (2009). Religious attendance in cross-national perspective: A multilevel analysis of 60 countries. American Journal of Sociology, 115(3): 863-895.

Stegmueller, D. (2013). How many countries do you need for multilevel modeling? A comparison of Bayesian and frequentist approaches. American Journal of Political Science, 57, 748–761.

Weldon, S. (2006). The institutional context of tolerance for ethnic minorities: A comparative, multilevel analysis of Western Europe. American Journal of Political Science, 40(2): 331-349.

Appendices

Appendix 1a. Structure of the Random intercept models used in the simulation study

Model specifications	**Values/categories**	**Experimental conditions**
Level 1		
Sample size	1000	-
Variables: X_1- X_5	Standardized effect size = .25, mean = 0, Variance = 1	-
Correlation between variables	0	-
Level 2		
Sample size	15, 20, 30, 40	4
Variables:		
Z_1	Standardized effect size = .10; .25, .50, mean= 0, variance= .50	3
Z_2-Z_5	Standardized effect size = .25, var. = .50	-
Number of variables in the model	1 (Z_1), 2 (Z_1, Z_2), …., 5(Z_1, Z_2, Z_3, Z_4, Z_5))	5
Correlation between variables	.25	-
Intraclass correlation	.10	-
Estimator	Maximum Likelihood, Bayes	2

Appendix 1b. Structure of the Random slope and cross-level interaction models

Model specifications	**Values/categories**	**Experimental conditions**
Level 1		
Sample size	1000	-
Variables		
X_1	Standardized effect size = .25	-
X_2-X_5	Standardized effect size = .25, variance = 1	-
Correlation between variables	0	-
Level 2		
Sample size	15, 20, 30, 40	4
Variables:		-
Z_1	Standardized effect size = .25, var. = .50	-
Cross-level interaction X_1Z_1	Standardized effect size = 0, .25, .50	3
Random intercept	0	-
Random slope	depends on X_1Z_1	-
Random slope variance (RSV)	depends on X_1Z_1	-
Random intercept variance (RIS)	.111	-
Correlation between RIV and RSV	.25	-
Intraclass correlation	.10	-
Estimator	Maximum Likelihood, Bayes	2

Appendix 2a. Parameter bias for the country-level effect – various country-level sample sizes, effect sizes, estimation procedures, and number of level-2 variables – 120,000 replications

	Low effect size				**Medium effect size**				**Large effect size**			
Sample size	15	20	30	40	15	20	30	40	15	20	30	40
M1: One country-level variable												
Maximum Likelihood	-.10	-.11	**-.02**	**-.01**	**-.04**	**-.04**	**-.01**	**.00**	**-.02**	**-.02**	**.00**	**.00**
Bayesian	-.07	**-.02**	**.03**	**-.01**	**-.03**	**-.01**	**.01**	**-.01**	**-.01**	**.00**	**.01**	**.00**
M2: Two country-level variables												
Maximum Likelihood	**-.02**	.07	**.02**	**.02**	**-.01**	**.03**	**.01**	**.01**	**.00**	**.01**	**.00**	**.00**
Bayesian	.06	.14	**.04**	**.01**	**.02**	**.05**	**.02**	**.00**	**.01**	**.02**	**.01**	**.00**
M3: Three country-level variables												
Maximum Likelihood	**.03**	.06	.09	.06	**.01**	**.02**	**.04**	**.02**	**.00**	**.01**	**.02**	**.01**
Bayesian	-.31	-.25	-.14	-.18	-.12	-.09	-.06	-.07	**-.05**	**-.04**	**-.03**	**-.03**
M4: Four country-level variables												
Maximum Likelihood	**.03**	**-.01**	**.01**	**.03**	**.01**	**.00**	**.00**	**.01**	**.00**	**.00**	**.00**	**.00**
Bayesian	-.06	**-.02**	**.04**	**-.01**	**-.02**	**-.01**	**.01**	**.00**	**-.01**	**.00**	**.01**	**.00**
M5: Five country-level variables												
Maximum Likelihood	**-.03**	-.08	**-.04**	**-.02**	**-.01**	**-.03**	**-.01**	**-.01**	**.00**	**-.01**	**.00**	**.00**
Bayesian	**-.03**	**-.05**	**-.01**	**-.03**	**-.01**	**-.02**	**.00**	**-.01**	**.00**	**.00**	**.00**	**.00**
M6: Collapsed over model complexity condition												
Maximum Likelihood	**-.02**	**-.01**	**.01**	**.02**	**-.01**	**.00**	**.01**	**.01**	**.00**	**.00**	**.00**	**.00**
Bayesian	-.08	**-.04**	**-.01**	**-.04**	**-.03**	**-.02**	**.00**	**-.02**	**-.01**	**-.01**	**.00**	**-.01**

Appendix 2b. Standard error bias for the country-level effect –various country-level sample sizes, effect sizes, estimation procedures, and number of level-2 variables – 120,000 replications

	Low effect size				**Medium effect size**				**Large effect size**			
Sample size	15	20	30	40	15	20	30	40	15	20	30	40
M1: One country-level variable												
Maximum Likelihood	-.14	-.11	**-.07**	**-.07**	-.14	-.11	**-.07**	**-.07**	-.14	-.11	**-.07**	**-.07**
Bayesian	.17	**.10**	**.05**	**.01**	.17	**.10**	**.06**	**.01**	.17	.11	**.06**	**.01**
M2: Two country-level variables												
Maximum Likelihood	-.11	-.11	**-.07**	**-.05**	-.11	-.11	**-.07**	**-.05**	-.11	-.11	**-.07**	**-.05**
Bayesian	.19	.11	**.05**	**.01**	.19	.11	**.04**	**.01**	.19	.12	**.04**	**.01**
M3: Three country-level variables												
Maximum Likelihood	-.19	-.14	-.11	**-.05**	-.19	-.14	-.11	**-.05**	-.19	-.14	-.11	**-.05**
Bayesian	.17	**.10**	**.04**	**.01**	.17	**.09**	**.04**	**.02**	.17	**.09**	**.04**	**.02**
M4: Four country-level variables												
Maximum Likelihood	-.26	-.16	**-.07**	**-.07**	-.26	-.16	**-.07**	**-.07**	-.26	-.16	**-.07**	**-.07**
Bayesian	.22	.13	**.05**	**.03**	.22	.14	**.05**	**.03**	.23	.13	**.05**	**.03**
M5: Five country-level variables												
Maximum Likelihood	-.28	-.20	-.15	**-.08**	-.28	-.20	-.15	**-.08**	-.28	-.20	-.15	**-.08**
Bayesian	.20	.14	**.02**	**.03**	.19	.14	**.02**	**.03**	.22	.14	**.02**	**.03**
M6: Collapsed over model complexity condition												
Maximum Likelihood	-.20	-.15	**-.09**	**-.07**	-.20	-.15	**-.09**	**-.07**	-.20	-.14	**-.10**	**-.07**
Bayesian	.19	.12	**.04**	**.02**	.19	.12	**.04**	**.02**	.20	.12	**.04**	**.00**

Appendix 2c. Coverage for the country-level effect –various country-level sample sizes, effect sizes, estimation procedures, and number of level-2 variables – 120,000 replications

	Low effect size				**Medium effect size**				**Large effect size**			
Sample size	15	20	30	40	15	20	30	40	15	20	30	40
One country-level variable												
Maximum Likelihood	.88	.91	.91	.92	.88	.91	.91	.92	.88	.91	.91	.92
Bayesian	**.96**	**.96**	**.95**	.94	**.96**	**.95**	**.95**	.94	**.97**	**.95**	**.95**	**.95**
Two country-level variables												
Maximum Likelihood	.90	.91	.92	.94	.90	.91	.92	.94	.90	.91	.92	**.96**
Bayesian	**.96**	**.96**	**.95**	.94	**.96**	**.96**	**.95**	.94	**.96**	**.95**	.94	.93
Three country-level variables												
Maximum Likelihood	.89	.89	.91	.94	.89	.89	.91	.94	.88	.89	.91	.93
Bayesian	**.97**	**.96**	.94	.94	**.97**	**.96**	**.95**	.94	**.97**	**.96**	.94	.94
Four country-level variables												
Maximum Likelihood	.85	.88	.93	.93	.85	.88	.93	.93	.85	.88	.93	.93
Bayesian	**.97**	**.96**	**.96**	**.95**	**.97**	**.96**	**.96**	**.95**	**.97**	**.96**	**.96**	**.95**
Five country-level variables												
Maximum Likelihood	.83	.87	.90	.93	.83	.87	.90	.93	.83	.87	.90	.92
Bayesian	**.97**	**.96**	.94	**.95**	**.97**	**.96**	.94	**.95**	**.97**	**.96**	.94	**.95**
Collapsed over model complexity condition												
Maximum Likelihood	.87	.89	.91	.93	.87	.89	.91	.93	.87	.89	.91	.93
Bayesian	**.97**	**.96**	**.95**	.94	**.97**	**.96**	**.95**	.94	**.97**	**.96**	**.95**	.94

Appendix 2d. Statistical power for detecting a country-level effect – various country-level sample sizes, effect sizes, estimation procedures, and number of level-2 variables – 120,000 replications

	Low effect size				**Medium effect size**				**Large effect size**			
Sample size	15	20	30	40	15	20	30	40	15	20	30	40
M1: One country-level variable												
Maximum Likelihood	.13	.11	.11	.12	.23	.24	.31	.38	.60	.70	**.86**	**.93**
Bayesian	.05	.05	.09	.10	.12	.15	.26	.32	.42	.58	.77	**.90**
M2: Two country-level variable												
Maximum Likelihood	.12	.12	.11	.13	.24	.28	.34	.40	.65	.78	.89	.96
Bayesian	.04	.06	.09	.10	.10	.16	.25	.33	.44	.62	.84	.92
M3: Three country-level variables												
Maximum Likelihood	.14	.14	.13	.13	.27	.28	.37	.44	.68	.78	**.90**	**.97**
Bayesian	.04	.05	.07	.09	.08	.14	.23	.30	.37	.56	**.82**	**.92**
M4: Four country-level variables												
Maximum Likelihood	.17	.15	.13	.15	.34	.34	.42	.51	.86	.94	**.99**	**1**
Bayesian	.03	.06	.08	.10	.10	.17	.31	.42	.57	.81	**.98**	**1**
M5: Five country-level variables												
Maximum Likelihood	.20	.17	.17	.17	.40	.44	.55	.64	1	1	1	**1**
Bayesian	.04	.05	.09	.10	.14	.20	.41	.56	**.96**	**1**	**1**	**1**
M6: Collapsed over model complexity condition												
Maximum Likelihood	.13	.14	.13	.14	.30	.31	.40	.47	.76	.84	.93	**.97**
Bayesian	.04	.05	.08	.10	.11	.16	.29	.38	.55	.71	**.88**	**.95**

Appendix 3a. Parameter bias for the country-level effect and a cross-level interaction effect – various country-level sample sizes, effect sizes, estimation procedures, and model complexity – 24,000 replications

	Z_1				X_1Z_1			
Sample size	15	20	30	40	15	20	30	40
M1: Random slope model								
Maximum likelihood	.06	**.03**	**.01**	**-.01**	-	-	-	-
Bayesian	**-.02**	**.01**	**.01**	**.03**	-	-	-	-
M2: Medium cross-level interaction								
Maximum likelihood	.06	**.05**	**.02**	**.00**	**.00**	**.04**	**.02**	**.02**
Bayesian	**-.03**	**.01**	**.00**	**.03**	**-.01**	**-.01**	**-.01**	**.01**
M3: Large cross-level interaction								
Maximum likelihood	.06	**.05**	**.02**	**.00**	**.00**	**.02**	**.01**	**.01**
Bayesian	**-.03**	**.00**	**.00**	**.04**	**-.01**	**.00**	**.00**	**.00**

Appendix 3b. Standard error bias for the country-level effect and a cross-level interaction effect – various country-level sample sizes, effect sizes, estimation procedures, and model complexity – 24,000 replications

	Z_1				X_1Z_1			
Sample size	15	20	30	40	15	20	30	40
M1: Random slope model								
Maximum likelihood	-.13	**-.09**	**-.06**	**-.06**	-	-	-	-
Bayesian	.22	**.11**	**.04**	**.04**	-	-	-	-
M2: Medium Cross-level interaction								
Maximum likelihood	**-.07**	**-.06**	**-.05**	**-.04**	**-.12**	**-.06**	**-.05**	**-.03**
Bayesian	.36	.19	**.07**	**.04**	.36	.25	.14	**.07**
M3: Large Cross-level interaction								
Maximum likelihood	**-.07**	**-.06**	**-.05**	**-.04**	**-.12**	**-.06**	**-.06**	**-.03**
Bayesian	.36	.19	**.07**	**.05**	.37	.24	.13	**.08**

Appendix 3c. Coverage for the country-level effect and a cross-level interaction effect – various country-level sample sizes, effect sizes, estimation procedures, and model complexity – 24,000 replications

	Z_1				X_1Z_1			
Sample size	15	20	30	40	15	20	30	40
M1: Random slope model								
Maximum likelihood	.90	.92	.92	.94	-	-	-	-
Bayesian	**.98**	.92	.92	.94	-	-	-	-
M2: Medium Cross-level interaction								
Maximum likelihood	.91	.92	.92	.94	.89	.94	.93	.94
Bayesian	**.98**	**.98**	**.97**	**.96**	**.98**	**.98**	**.96**	**.96**
M3: Large Cross-level interaction								
Maximum likelihood	.91	.92	.92	.94	.91	.93	.93	.94
Bayesian	**.98**	**.98**	**.96**	**.96**	**.98**	**.97**	**.96**	**.96**

Appendix 3a. Statistical power for detecting a country-level effect and a cross-level interaction effect – various country-level sample sizes, estimation procedures, and model complexity – 24,000 replications

	Z_1				X_1Z_1			
Sample size	15	20	30	40	15	20	30	40
M1: Random slope model								
Maximum likelihood	.27	.27	.33	.40	-	-	-	-
Bayesian	.10	.15	.27	.36	-	-	-	-
M2: Medium cross-level interaction								
Maximum likelihood	.24	.27	.31	.38	.20	.23	.28	.34
Bayesian	.07	.12	.24	.35	.06	.10	.18	.26
M3: Large cross-level interaction								
Maximum likelihood	.24	.26	.31	.38	.61	.74	**.86**	**.95**
Bayesian	.05	.12	.24	.33	.32	.49	.76	**.89**

Appendix 4. Statistical power for detecting a country-level effect equal to zero – various country-level sample sizes, estimator procedures and model complexity – 24,000 replications

Sample size	15	20	30	40
Random intercept model (Z_1=0)				
Maximum Likelihood	0.12	0.09	0.09	0.08
Bayesian	0.03	0.04	0.05	0.06
Random intercept, random slope model (Z_1=0)				
Maximum Likelihood	0.10	0.08	0.08	0.07
Bayesian	0.02	0.04	0.05	0.04
Cross-level interaction model (X_1Z_1 = 0)				
Maximum Likelihood	0.09	0.07	0.07	0.06
Bayesian	0.02	0.02	0.03	0.04