The Endogeneity Problem in Random Intercept Models: Are Most Published Results Likely False?

Conference Paper *in* Academy of Management Proceedings · August 2019

DOI: 10.5465/AMBPP.2019.18927abstract

CITATIONS
12

READS
2,104

3 authors:

John Antonakis
University of Lausanne
167 PUBLICATIONS **16,299** CITATIONS
SEE PROFILE

Nicolas Bastardoz
KU Leuven
38 PUBLICATIONS **1,247** CITATIONS
SEE PROFILE

Mikko Rönkkö
University of Jyväskylä
106 PUBLICATIONS **3,267** CITATIONS
SEE PROFILE

On ignoring the random effects assumption in multilevel models:

Review, critique, and recommendations

John Antonakis

University of Lausanne

john.antonakis@unil.ch

Nicolas Bastardoz

University of Zurich

nicolas.bastardoz@uzh.ch

Mikko Rönkkö

University of Jyväskylä

mikko.ronkko@jyu.fi

In press:

Special issue on New Approaches to Multilevel Methods and Statistics

Organizational Research Methods

Author Note

This research was supported in part by a grant from the Academy of Finland (grant 311309) to Mikko Rönkkö. We acknowledge the computational resources provided by the Aalto Science-IT project.

Author bios

John Antonakis is professor of organizational behavior at the Faculty of Business and Economics, of the University of Lausanne, Switzerland. His research is focused on applied methodological issues and causality, in addition to substantive organizational behavior topics like leadership and individual differences.

Nicolas Bastardoz is senior research associate at the chair of Human Resource Management and Leadership, of the University of Zurich, Switzerland. His research interests include charismatic leadership, followership, and quantitative research methods.

Mikko Rönkkö is associate professor of entrepreneurship at Jyväskylä University School of Business and Economics, Finland. His current research interests are statistical research methods and software entrepreneurship.

Abstract

Entities such as individuals, teams, or organizations can vary systematically from one another. Researchers typically model such data using multilevel models, assuming that the random effects are uncorrelated with the regressors. Violating this testable assumption, which is often ignored, creates an endogeneity problem thus preventing causal interpretations. Focusing on two-level models, we explain how researchers can avoid this problem by including cluster means of the Level 1 explanatory variables as controls; we explain this point conceptually and with a large scale simulation. We further show why the common practice of centering the predictor variables is mostly unnecessary. Moreover, to examine the state of the science, we reviewed 204 randomly drawn articles from macro and micro organizational science and applied psychology journals, finding that only 106 articles—with a slightly higher proportion from macro-oriented fields—properly deal with the random effects assumption. Alarmingly, most models also failed on the usual exogeneity requirement of the regressors, leaving only 25 mostly macro-level articles that potentially reported trustworthy multilevel estimates. We offer a set of practical recommendations for researchers to model multilevel data appropriately.

Keywords: random effects, fixed effects, multilevel, HLM, endogeneity, centering

On ignoring the random effects assumption in multilevel models: Review, critique, and recommendations

Researchers often apply data that varies on multiple levels. The two main traditions for working with such hierarchical data are (a) regression models for panel data as used in the economics (Wooldridge, 2002, Chapter 11; Wooldridge, 2013, Chapter 14) and sociology (Allison, 2009; Halaby, 2004) and (b) multilevel models as typically used in the education, management, and psychology literatures (e.g., Helson, Jones, & Kwan, 2002; Hofmann, 1997; Lee, 2000). Whereas there is no shortage on methodological research addressing these topics, the literature on hierarchical data is rather technical and strongly fragmented so that texts addressing econometrics techniques for panel data rarely present multilevel modeling as an alternative solution and vice versa (cf. McNeish & Kelley, 2018). The fragmented literature can make it difficult for researchers to develop a solid understanding of the main issues that must be addressed and the different ways of doing so.

Focusing on two-level models—which are the most common as our review will indicate—we integrate the two main traditions to explain how they address the challenges in modeling multilevel data. We make two key contributions. First, we show ways to appropriately estimate models for hierarchical data, highlighting a major threat to estimate validity, endogeneity, a confounding that can stem from various sources (Antonakis, Bendahan, Jacquart, & Lalive, 2014). In the case of multilevel models, endogeneity can arise when the assumption that the Level 2 error term is uncorrelated with the Level 1 regressors is violated, which renders the coefficients of the Level 1 regressors causally uninterpretable. In the case of this article, we will show how endogeneity is introduced into the model from failure to correctly model the unobserved variation due to the hierarchical structure of the data and where the researcher is interested in estimating what is called the “within” effect (i.e., the effect of a Level 1 regressor on

a Level 1 outcome); it is this effect, which is commonly of interest and relevant for causal interpretation. We show the source of the endogeneity problem, how to deal with the issue using straightforward procedures (i.e., including the cluster means of all Level 1 regressors), demonstrate the workings of various estimation methods with simulated data, and then using an extensive Monte Carlo simulation we help researchers understand the hazards at hand in an intuitive way. Appendix A provides links to video materials explaining these problems.

Our second contribution will be to take stock of the literature by describing what researchers typically do when estimating such models. That is, we will scrutinize the literature from the management and applied psychology fields to examine the extent to which researchers estimate multilevel models appropriately and whether the results of research published using such models can be viewed as trustworthy. Our review indicates that the critical assumptions in multilevel analysis—that the unobserved random effects are uncorrelated with the Level 1 regressors (i.e., Raudenbush & Bryk, 2002, p. 255)—are often ignored; in doing so, the results of such studies are unnecessarily exposed to conditions that are known to compromise estimate validity. Our article should thus serve as a warning siren for researchers estimating multilevel models. We will conclude by making easy-to-follow recommendations to authors, reviewers, and editors to help ensure the validity of published multilevel research.

Unobserved heterogeneity in hierarchical datasets

We begin by introducing the basic random-intercept model, which serves as a starting point for more complex multilevel models (Bliese & Ployhart, 2002; Holcomb, Combs, Sirmon, & Sexton, 2010). The random intercept model is a special case of a basic multilevel model, shown in Eq. 1 below, where a lower level unit i is nested in a higher-level unit j often referred to as a "cluster" (note, we will extend this model later to accommodate regressors at Level 2 as well). In micro-level research, a common scenario is nesting of subordinate i under leader j; in such a case

differences in team leaders' skills may allow members of some teams to perform systematically better than members of other teams. In macro-level research the setting could be firm i nested in industry j; here, firms in some industries may have systematically different profitability than do firms in other industries because of different asset requirements between the industries. Nesting also occurs with panel data for instance, wherein observations in year i are nested in firm j.

In the basic case, we are interested in how the subordinate (or firm) level characteristic affects the subordinate (or firm) level outcomes taking the nested structure of the data into consideration. Using the basic notation of Raudenbush and Bryk (2002), we have the following multilevel model:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij} \quad \} \; Level\ 1\ equation \qquad Eq.\ 1a$$

$$\left.\begin{array}{l} \beta_{0j} = \gamma_{00} + u_{0j} \\ \beta_{1j} = \gamma_{10} + u_{1j} \end{array}\right\} Level\ 2\ equations \qquad \begin{array}{l} Eq.\ 1b \\ Eq.\ 1c \end{array}$$

The Level 1 equation states that the expected value of the dependent variable *y* depends linearly on the observed *x* value; the Level 2 equations state that the linear dependency shown in the Level 1 equation can vary between the Level 2 units so that each unit is allowed to have a unique intercept and slope in the Level 1 equation. That we allow each cluster (e.g. leader, industry) to have a separate intercept and slope is indicated by the subscript *j* for the regression coefficients β in the Level 1 equation. The term e_{ij}—variation of *y* not due to *x*—is assumed to be uncorrelated with *x*; this assumption is a standard one made in regression (e.g. Wooldridge, 2013, assumption MLR.4). In the Level 2 equations, the γ's are the mean values of the slope and intercept over all clusters, and the u terms define how the slopes and intercept vary between cluster. The regression coefficients β and γ are fixed parameters having specific values, shared between all observations. The u and e terms are unobserved and modeled as random effects;

whereas we assume that each Level 1 or Level 2 unit has a specific value, these values are not estimated directly but their distribution is—assuming the effect is normally distributed, this reduces to estimating its variance. The u and e terms are assumed to be independent of each other.

Eq. 1a can be re-expressed by substituting β_{0j} and β_{1j} in the y_{ij} equation:

$$y_{ij} = \underbrace{\gamma_{00} + \gamma_{10}x_{ij}}_{fixed\ part} + \underbrace{x_{ij}u_{1j} + u_{0j} + e_{ij}}_{random\ part} \qquad Eq.\ 2$$

The above model representation is called a mixed effects model because it is composed of two parts: (a) The fixed part contains only fixed coefficients and gives the regression line, and (b) the random part contains only random effect terms indicating how observations vary around the line.

Having introduced the general model, we will now focus on the special case of a random intercept model where we assume that slopes β_{1j} do not vary between clusters (i.e., $u_{1j} = 0$); we thus drop $x_{ij}u_{1j}$ from the equation and for further simplicity, we remove the superfluous subscript $_0$, obtaining a simpler equation that we label as "RE model". We also present another model variant, labeled as "FE model", where the cluster specific intercepts are modeled in the fixed part. We use the labels RE and FE exclusively to refer to these two models to avoid the general confusion around the statistical concepts of fixed and random effects; fixed effects are parameters for which specific values are estimated and random effects are parameters for which we estimate a distribution. In the HLM literature researchers typically refer to the slope being "fixed" when a slope does not have a random effect (e.g. u_{1j} is omitted from Eq. 1c) and use the term "fixed effect" in this context. For instance, Hofmann (1997, p. 729) states: "Fixed effects are parameter estimates that do not vary across groups" (see also p. 77 in Raudenbush & Bryk, 2002). However, in econometrics text the terms "fixed effect" and "random effect" are commonly

used in the context of explaining the difference between the RE model and FE model (Wooldridge, 2013, Chapter 13). This difference in presentation and emphasis can lead to confusion when talking across literatures (McNeish & Kelly, 2018; see also Clark & Linzer, 2015). Suffice it to say that researchers should pay close attention to what the author means when using the term "fixed effect" (Wooldridge, 2002, p. 251-252). The two models are shown below:

RE model:
$$y_{ij} = \underbrace{\gamma_0 + \gamma_1 x_{ij}}_{fixed\ part} + \underbrace{u_j + e_{ij}}_{random\ part} \qquad Eq.\ 3a$$

FE model:
$$y_{ij} = \underbrace{\gamma_0 + \gamma_1 x_{ij} + a_j}_{fixed\ part} + \underbrace{e_{ij}}_{random\ part} \qquad Eq.\ 3b$$

In the RE model, the u_j term that represents the unobserved between cluster variation (e.g. differences between leaders/industries that are not explicitly measured) is a random intercept, hence the term "random intercept model." However, the between cluster variation need not be modeled in the random part of the model; it could equally well be assigned to the fixed part of the model, thus producing the FE model in Eq. 3b. Here we replace the random effect u_j with the fixed effect a_j. The difference between these approaches is that in the FE approach, cluster specific values are estimated for a_j (e.g., by using dummy variables) whereas in the RE approach just the variance but not case values of u_j are estimated. Outside any particular model, the phenomenon that some Level 2 units are higher on the dependent variable than others is referred to in econometrics as "unobserved heterogeneity" (Wooldridge, 2013, Chapter 13), and has been given various other labels in organizational research such as "stable variance" (Guo, 2017).

As is evident above, whereas both u_j and a_j represent the same feature in the data, they do so in a fundamentally different way; failure to understand this difference is the key to the

endogeneity problem in modeling multilevel data. A key assumption in both the RE model (Eq. 3a) and the FE model (Eq. 3b) is that the random part of the model is uncorrelated with the regressors. The random part of the RE model has, in addition to the e_{ij} term, the u_j term too. Thus, compared to the FE model, the RE model has an additional assumption that $cov(x_{ij}, u_j) = 0$, which is referred to as the *random effects assumption* in econometrics. Failure of this assumption compromises the accuracy of estimates. To be more precise, by accuracy we mean (a) estimate consistency (i.e., whether coefficient estimates approach the correct values when sample size increases), and (b) estimate bias (i.e., whether the mean estimates of repeated sampling will converge to the correct estimates).

Examples can clarify the meaning of the random effects assumption. Suppose we wish to examine how leaders affect subordinate performance. It is likely that something in the leader (e.g., intelligence) correlates with subordinate-level characteristics; smarter leaders might assign tasks differently to conscientious individuals, give them more autonomy, monitor them less, and so forth; consequently, the subordinates are better performing. It is possible too that something in the leader (e.g., intelligence) directly drives subordinate performance; for instance, smarter leaders, by virtue of their expertise may provide greater clarity about how to better accomplish tasks. Thus, leader intelligence may have an effect on y_{ij} and also correlate with x_{ij}. Failure to measure and explicitly model intelligence will thus engender endogeneity bias because leader intelligence is an omitted variable. In reality, intelligence is not the only Level 2 factor that may contribute to the systematic differences. The sex of the leader may matter; so too may age, experience, personality, looks, height, testosterone level, or what have you (Antonakis, 2011).

As an example in macro-level work, firms nested in different countries may face different legal requirements. Thus, country legal origin may correlate with firm level characteristics;

however, legal origin could also drive firm outcomes (see La Porta, Lopez-De-Silanes, & Shleifer, 2008). So may climatic factors or location factors like being in a landlocked country (see Sachs, 2003). There are many country-level differences that may matter; but, which of them matter? It is very hard to know because we may miss some important, but unobservable factors. To the extent that an effect of the Level 2 variable that matters (i.e., constant at Level 2) is not included in the model introduces endogeneity bias because $cov(x_{ij}, u_j) \neq 0$. Thus, failing to include all pertinent variables at the leader (for the micro case) or country (for the macro case) level will introduce endogeneity into the model leading to biased and inconsistent estimates (refer to Appendix B for a formal explanation).

Modeling approaches in organizational research

After introducing the general problem and the two dominant modeling approaches for unobserved Level 2 differences, RE and FE models, we now turn to specific techniques that have been used in past organizational research. We do so for two reasons. First, as explained earlier, the literature on these techniques is fragmented (see also McNeish & Kelley, 2018) and thus a review that integrates the techniques between the multilevel modeling and econometrics literatures can help in making more informed choices of modeling techniques. Second, researchers seem to choose their techniques from one of these literatures without consulting the other. Therefore, providing specific examples of the analysis techniques that researchers use can be helpful when the researchers are trying to understand the bigger picture that we discuss in the article. Thus, we will explain the various techniques used for addressing unobserved heterogeneity and whether the techniques fall into the RE or FE models. We also shed light on the key equivalencies or differences between the techniques, which are summarized in Table 1.

[Insert Table 1 here].

FE modeling approaches

We first discuss the FE approach. The two general modeling strategies are (a) to include a set of dummy variables in the model, which explicitly estimate a separate intercept for each Level 2 unit in the data, and (b) transformations that eliminate the between-cluster differences from the data before estimation. The key idea in these techniques is to estimate a specific value of a_j (a fixed effect) for each cluster and then eliminate the effect of these values from the data. Whereas both these approaches do so differently, they both accomplish the same result: They eliminate all higher level effects from the data allowing estimating the within effect γ_1 consistently.

Dummy variable regression

Using dummy variables for each Level 2 unit as controls in a model is perhaps the easiest way to understand the FE modeling approach. This approach is shown in the following equation:

$$y_{ij} = \gamma_0 + \gamma_1 x_{ij} + \sum_{k=2}^{n} \delta_k D_j + e_{ij} \qquad \textit{Eq. 4}$$

In the above, the a_j term is replaced by a set of dummy variables *D*, indicating the *n-1* Level 2 units leaving out the first unit as the baseline. In practice, estimating a model with dummy variables is inconvenient because the size of the estimated model can make the model computationally difficult and produces extensive printout from statistical software. Because the specific group differences are rarely of direct interest, researchers generally apply another, simpler estimation approach that produces the same results.

Cluster-mean centering and GLS fixed effects estimator

Centering a variable involves subtracting the mean from the variable, and in the case of cluster-mean centering, the mean is subtracted for each cluster separately. Centering is an intensely discussed topic in the literature on multilevel models (Enders & Tofighi, 2007; Hofmann &

Gavin, 1998; Kreft, De Leeuw, & Aiken, 1995; Raudenbush & Bryk, 2002) and cluster-mean centering has a prominent role in in the econometrics literature as well. In fact, if one uses cluster-mean centering, the fixed part of the basic two level example model can be estimated consistently with ordinary least squares (OLS) regression. This approach is referred to as *generalized least squares* (*GLS) fixed effects estimator* or *GLS within estimator* in econometrics (Greene, 2012, sec. 9.3; Wooldridge, 2013, p. 485). Thus, GLS fixed effects estimates the following model by OLS:

$$(y_{ij} - \bar{y}_j) = \gamma_1(x_{ij} - \bar{x}_j) + e_{ij} \qquad \textit{Eq. 5}$$

A key feature in Equation 5 is that all independent variables and the dependent variable are cluster-mean centered: The centering of the predictors guarantees that the cluster-means of the fitted values are zero and thus all between-cluster differences due to the observed predictors are eliminated. Cluster-mean centering the dependent variable will then eliminate all unobserved differences, a_j, between the clusters and thus produces the same estimates as the dummy variable regression. However, this latter transformation requires adjustments to the standard errors (Wooldridge, 2002 pp. 269-272), which is sometimes overlooked in the multilevel modeling literature (e.g. Raudenbush & Bryk, 2002, pp. 135–137).

Whereas the FE approaches are convenient to apply because they do not require the restrictive assumption that u_j is uncorrelated with the observed regressors, these approaches have two limitations: (a) They are inefficient when the random effects assumption holds (i.e., estimates are less precise, though this issue is less relevant in large samples), and most notably (b) cannot be used when the interest is on the effects of Level 2 variables—which do not vary within cluster—on *y* because the Level 2 variables are perfectly collinear with the data that account for the FEs.

The latter restriction is commonly used as a justification to apply the RE approach, warranting an explanation of the collinearity problem. The problem is more apparent in the techniques that cluster-mean center: If a variable only varies between clusters, then cluster-mean centering removes all this variation leaving no variation in a variable. Estimating an effect of a variable that does not vary is impossible (Wooldridge, 2013, MLR.3 assumption). In the dummy variable regression case, the mechanism is less obvious but still present. Here, the problem is that the set of dummies, by design, explain all between cluster variation and thus their effect completely overlaps with any Level 2 variables. The situation is similar to trying to estimate the effects of CEO gender and marital status on some outcome but only having a sample consisting of married men and unmarried women; because of complete overlap between gender and marital status, it is impossible to estimate their unique effects. To see this collinearity problem more clearly, refer to Table 3 to see how the Level 2 regressor, z_j is perfectly collinear to the dummy variables; in this case the Level 2 regressor is a binary variable, but the same problem would arise if it were continuous given that within cluster the Level 2 regressor is constant.

RE modeling approaches

The key idea of RE approaches is that instead of estimating specific values for each cluster, we estimate how much of the variance of the random part–assuming normal distribution of u_j and e_{ij}–is due to the unobserved Level 2 effect. In practice, this approach means estimating the variance of the Level 2 effect, u_j, or equivalently the intraclass correlation of the random part of the model $u_j + e_{ij}$. This procedure can be done either by calculating the variance of the Level 2 effect first, and then (a) either transforming the data to eliminate the effect or (b) by adjusting the standard errors formula and possibly the estimation criterion to take the clustering effect into account, or (c) by estimating the full model simultaneously.

GLS and maximum likelihood (ML) estimation of RE model

The two principal approaches for estimating this model are the *GLS random effects estimator* as discussed in economics (Wooldridge, 2013, Chapter 14.2) and maximum likelihood estimation of the RE model in Eq. 3a under the assumption that u is normally distributed, discussed extensively in multilevel modeling literature (e.g., Raudenbush & Bryk, 2002; Hofmann, 1997). The economics approach applies OLS to transformed data as follows:

$$(y_{ij} - \lambda\bar{y}_j) = \gamma_0(1 - \lambda) + \gamma_1(x_{ij} - \lambda\bar{x}_j) + e_{ij} \qquad \textit{Eq. 6}$$

This approach applies quasi-mean centering, because cluster means are not subtracted directly, but are scaled by the λ parameter that depends on the estimated variance of the Level 2 effect. The GLS random effects estimates are thus a weighted average of OLS estimates and GLS fixed effects estimates and λ varies between 0 (GLS RE produces OLS estimates) and 1 (GLS RE produces GLS FE estimates). The basic idea of this estimation approach is that when the random effects assumption holds, the parameters can be consistently and unbiasedly estimated by using cluster means of all variables in what is usually called a "between regression". The assumption that the between and within effects are the same is also mentioned in the multilevel modeling literature, where the terms individualistic (Yammarino & Dansereau, 2011), atomistic, and ecological fallacy are used (Luke, 2004, pp. 5–6; Rabe-Hesketh, 2012, sec. 3.7; Raudenbush & Bryk, 2002, pp. 135–139). Thus, the GLS RE estimation approach can improve the efficiency of estimation over GLS FE by also taking the information from the between regression into account.

OLS or generalized estimating equations (GEE) and cluster robust standard errors

The third approach to estimate the RE model is to simply apply normal OLS regression analysis. This approach relies on the fact that consistency and unbiasedness of OLS regression requires that the random part of the model is uncorrelated with the fixed part, but does not require that the

random part is independent between observations. However, if the random part is not independent between observations, OLS regression is inefficient and GEE provides a more efficient alternative (for details see Ballinger, 2004; McNeish, Stapleton, & Silverman, 2017).

Regardless of which estimation approach is applied, the conventional standard errors are inconsistent and cluster robust variant that allows for arbitrary correlations within cluster must be applied (Angrist & Pischke, 2008, see Ch. 9); these standard errors are also useful when there is autocorrelation over time. Although this approach has an advantage in its simplicity and robustness, it can be less efficient than direct ML estimation of the RE model and does not provide an estimate of the unobserved between cluster variance.

The RE approaches, regardless which implementation is chosen, have the added advantage that they avoid the collinearity problem that prevents estimating of Level 2 effects in the FE approaches. Thus, these approaches allow estimating models such as the following:

$$y_{ij} = \gamma_0 + \gamma_1 x_{ij} + \gamma_2 z_j + u_j + e_{ij} \qquad Eq.\ 7$$

Where z_j is a Level 2 variable that does not vary within clusters. However, as mentioned before, it is critical to include all relevant Level 2 variables in the model else u_j will correlate with the Level 1 regressors rendering the estimator inconsistent and biased. Note, the Eq. 7 is a more general version of the basic multilevel equation because it now includes both Level 1 and Level 2 regressors (of course, such a model can accommodate multiple Level 1 and Level 2 regressors).

Correlated random effects (CRE) modeling approaches

The assumption that the random intercept is uncorrelated with the regressors, made in the RE modeling approach explained in the previous section, is restrictive, often violated, and as our review will indicate, often not properly considered by researchers. Fortunately, the assumption that u_j is uncorrelated with the Level 1 regressors is testable, and can be relaxed by adding

cluster means to the model. This approach originates from the work by Mundlak (1978) and produces what some econometricians refer to the correlated random effects approach (Wooldridge, 2013, pp. 497–499). Whereas the terminology is similar, the CRE approach should not be confused with the common practice of allowing a correlation between random intercepts and slopes. Instead, to specify a CRE model, one simply includes the cluster mean of the x_{ij} regressor in the model and estimates (Allison, 2009, pp. 23-27; Antonakis, Bendahan, Jacquart, & Lalive, 2010; Neuhaus & Kalbfleisch, 1998; Rabe-Hesketh & Skrondal, 2008, p. 119; Schunck, 2013):

CRE model:
$$y_{ij} = \gamma_0 + \gamma_1 x_{ij} + \gamma_{2(contextual)}\bar{x}_j + u_j + e_{ij} \qquad \textit{Eq. 8a}$$

$$y_{ij} = \gamma_0 + \gamma_1 (x_{ij} - \bar{x}_j) + \gamma_{2(between)}\bar{x}_j + u_j + e_{ij} \qquad \textit{Eq. 8b}$$

Equation 8a shows the original CRE approach and 8b shows a variant of the same model where the explanatory variable has been cluster-mean centered. Interestingly, regardless of which of the two specifications is applied, these modeling approaches have been shown to be equivalent to the FE approach for estimating γ_1 (Wooldridge, 2013, Chapter 14.2; Enders & Tofighi, 2007).

Whereas γ_1 always provides the within effect, the value and interpretation of γ_2 depends on whether the Level 1 variable x_{ij} is cluster-mean centered or not, indicated by the added subscripts in γ_2, producing either the between effect or the contextual effect. To understand the contextual effect, it is useful to contrast it against the within and the between effects. Using a micro example, the within effect tells how characteristics or actions of an individual affect individual level outcomes, for example how individuals' performance depends on individuals' intelligence. The within effect has clear policy implications because it answers the question of what one can do to improve one's performance. The contextual effect, however, tells how

characteristics or actions of other individuals in the same context affect individual level outcomes or, alternatively, how characteristics or actions of an individual affect the outcomes of others in the same context, or how mean of the characteristic in the context affects individual level outcomes. For example, how individuals' performance depends on the average intelligence of the team controlling for one's own intelligence level. The between effect is simply the sum of these two effects. For example, how the average performance of a team depends on the average intelligence of the team members. Which of these effects is the most relevant depends on the research question that one wants to answer. Table 2 shows examples of within and contextual effects.

[Insert Table 2 here]

The contextual effect also provides a third way to understand the random effects assumption; in addition to the (a) uncorrelated unobserved cluster effect and (b) the equivalence of the within and between effects, the random effects assumption also means that (c) all contextual effects are zero.

The CRE model essentially unifies the RE and FE models. As an advantage over both models, this modeling approach provides the estimates for both within and between or contextual effects simultaneously (Antonakis, et al., 2010; Certo, Withers, & Semadeni, 2017; Curran & Bauer, 2011; Enders & Tofighi, 2007; McNeish & Kelley, 2018). As an advantage over FE modeling, the CRE model allows for modeling effects of variables that are constant within cluster (i.e., between or contextual effects); that is, one can add z_j variables to the models in Equations 8. As an advantage over the RE approach, it does not make the often-unrealistic assumption that the unobserved Level 2 term is uncorrelated with all regressors. As a final advantage, the CRE models makes it easier to transition between FE and RE and can test if unobserved Level 2

effects are correlated with the explanatory variables and if so, which ones. Moreover, whether or not there is endogeneity due to u_j, the within effect is consistently estimated as long as the assumption of the exogeneity of the regressors holds. However, given that the Level 1 estimates from the CRE approach are identical to the FE estimates (Wooldridge, 2013, Chapter 14.2; Enders & Tofighi, 2007), the CRE approach is less efficient than the RE approach if the random effects assumption holds (in large samples this issue is less relevant).

To better understand the data structure for estimating multilevel models, we show in Table 3 how some hypothetical data are set-up in the "long" format, where each repeated Level 1 observation is stored vertically in a new row of data (in contrast to the "wide" format where each repeated observation would be a new column variable), for the various approaches we discussed. This dataset includes a Level 1 predictor *x*, a Level 2 predictor, *z,* dummy variables to identify the Level 2 clusters, cluster means of *x*, as well as the key centering approaches.

[insert Table 3 here]

Empirical tests for the random effects assumption

Although the choice of a modeling approach can be justified from theory, empirical tests also exist to guide researchers and should be used as the final arbiters; however, these tests are not well known in the management and applied psychology literatures. Worse, some canonical sources have even suggested—incorrectly—that the decision on whether to apply cluster-mean centering "cannot be based on statistical evidence" (Enders & Tofighi, 2007, p. 135) or that the choice "can only be made on a theoretical basis" (Kreft, et al., 1995). The centering choice taken will determine if the within effect is consistently estimated. It is thus critical that researchers understand how to use the relevant tests, which are summarized in Table 4 and explained next.

[Insert Table 4 here]

Hausman test

The best-known test to examine the random effects assumption is the Hausman (1978) test, which can be used to compare an estimator that is assumed to be consistent (FE) against an efficient estimator (RE). The basic version of the Hausman test is:

$$H = \frac{(\hat{\gamma}_{CONSISTENT} - \hat{\gamma}_{EFFICIENT})^2}{SE(\hat{\gamma}_{CONSISTENT})^2 - SE(\hat{\gamma}_{EFFICIENT})^2} \qquad Eq.\ 9$$

The logic of the test is that if both estimators are consistent, then in large samples the two estimates should be very close to one another. The test statistic thus compares the two estimates, and if their difference is greater than what can be expected due to sampling error (quantified by the standard errors [SE] in the denominator), then the null hypothesis that both estimators are consistent is rejected. In practice, a statistically significant result suggests that the constraint that u_j is unrelated to x_{ij} and y_{ij} does not hold and the efficient RE estimator must be rejected.

The Hausman test has a few disadvantages: (1) the justification of the test relies on large sample results, (2) the test can lead to computational difficulties due to non-positive definite denominator, and (3) the test is only valid if using conventional standard errors; if robust standard errors are used, the general Wald test explained later can be applied instead (Wooldridge, 2002, pp. 290-291). Moreover, the Hausman test does not directly test the random effects assumption, but is a general test that can be applied to compare any estimators, where one estimator is assumed to be consistent and the other is possibly inconsistent. There are two tests, however, that directly examine the random effects assumption and both involve the use of the CRE model.

Likelihood ratio test

The likelihood ratio test can be used for comparing nested models. Two models are nested if they are fitted to the same data and one can be expressed as a constrained version of another. This test

applies to maximum likelihood estimates and therefore cannot be used when GLS estimation is used, and hence cannot be used to compare the FE model against the RE model. However, it can be used for comparing the maximum likelihood estimates of a RE model against the CRE model. A statistically significant χ^2 value indicates that the model with cluster means (the less constrained CRE model) fits the data statistically significantly better than does the RE model and thus implies rejection of the random effects assumption.

***F* test or Wald test**

When nested model comparisons are not possible, the *F* (or Wald χ^2) tests can be used for testing multiple parameter constraints after estimation, whether using OLS, GLS, or ML estimation. This technique involves estimating a CRE model and then performing a post estimation *F* test of the null hypothesis that all contextual effects (i.e., γ_2 in Eq. 8a) are zero. The Wald test is essentially a generalization of the *z* test to multiple parameters and is very versatile because it can be performed regardless how the variance was estimated (i.e., robust or cluster-robust standard errors), and is thus often recommended as an alternative to the Hausman test when its assumptions fail (e.g., Wooldridge, 2002, pp. 290-291). Equivalently, instead of testing the effects of added cluster means, cluster-mean centered versions of variables can be added along with the original variables (Arellano, 1993).

Demonstration using generated data

Suppose we have a sample of 500 leaders, each supervising a team of 10 individuals, or to use a macro example, 500 firms, each observed over 10 year. Our sample is fairly large so that the consistency of the estimators "kicks in" and sampling error would influence the results only minimally. We generated the data using the following model:

$$y_{ij} = .50x_{ij} - .50z_j + u_j + e_{ij} \qquad \textit{Eq. 10}$$

where u_j is correlated with the outcome as well as with the Level 1 (i.e., x_{ij}) and Level 2 (i.e., z_j) regressor. Whereas the endogeneity of the Level 1 variable with respect to the unobserved cluster effect can be handled with the techniques discussed in this article, endogeneity due to other reasons in the Level 1 or Level 2 estimates can only be eliminated by using instrumental variables (a discussion on instrumental variables is beyond the scope of the paper; readers should refer to more specialized literature to see how this estimation procedure can be used: Angrist & Pischke, 2014; Antonakis, et al., 2010; Bascle, 2008; Bollen, 2012; Gennetian, Magnuson, & Morris, 2008; Larcker & Rusticus, 2010).

Then we compare how the different modeling procedures affect parameter estimates. The data generating code in Stata and R can be found in the Appendix C; the R code uses plm (Croissant & Millo, 2008), lme4 (Bates, Mächler, Bolker, & Walker, 2015), and clubSandwich packages (Pustejovsky & Tipton, 2018). We estimate the models using Stata 15 (StataCorp, 2017) using cluster robust standard errors when required.

[insert Table 5 here]

The demonstration begins uneventfully, as shown in Table 5. As expected, the pooled OLS model that ignores the unobserved cluster effect produces very erroneous estimates (model 1); the coefficient of x = 2.97 and z = -.39 corresponding to rather large estimation errors. The OLS estimator with dummy variables for all leaders (model 2) provides a much better estimate for the coefficient of x =.51 and the GLS FE (model 3) provides the same estimate as does OLS using cluster means as controls (model 4). Using OLS and controlling the cluster means, including z makes no difference to the estimate for the within effect (model 5); the effect of z = -.47 is only slightly erroneous.

The rest of the results are more intriguing. First, using GLS RE where we include the cluster means (model 6), we get the same estimate as those of the previous model; though this approach has the advantage of estimating the variance components (i.e., the variances of e_{ij} and u_j from Eq. 11). Because we use the cluster robust standard errors, the Hausman test cannot be applied and we use the Wald test for $\bar{x}_j = 0$ instead. The test is significant giving $\chi^2(1) =$ 3,514.10, $p < .001$ (this value is the square of *t*-statistic for $\bar{x}_j$, i.e., 59.28^2 approximately), indicating that the random effects assumption does not hold and $\bar{x}_j$ must be included in the model. We get very similar estimates for the CRE model using maximum likelihood estimation (model 7); the likelihood ratio test comparing the model with and without the cluster means is significant, $\chi^2(1) = 1130.08$, $p < .001$ again suggesting that the random effects assumption must be rejected. Note that models 4 to 7 estimate the contextual effect (i.e., see Eq. 8a) as captured by the coefficient of $\bar{x}_j$. The next model (model 8) estimates the between effect (i.e., see Eq. 8b) as captured by the coefficient of $\bar{x}_j = 5.09$.

Things go haywire from here on! The GLS RE estimator that omits the cluster means (model 9) gives a very bad estimate and the coefficient of *x* is 1.23. The coefficient of *z* is -.33. After estimating a RE model, we test the RE assumption, this time using the *xtoverid* command (Schaffer & Stillman, 2006) that provides a convenient way of doing the Wald test, which again strongly rejects the constraint that the RE approach uses ($\chi^2 = 3{,}514.10$, $p < .001$). Cluster-mean centering *x* (model 10) provides the correct estimate for the within effect (i.e., the coefficient of $x - \bar{x}_j = .51$); however, the estimate for *z* is now rather off (i.e., -.28 instead of -.50). The ML RE estimator that uses the grand-mean centered data (model 11) does not provide any advantage with respect to using the original *x* (model 9), giving very poor estimates (and the constraint that the cluster means are zero is, of course, rejected).

As is evident from the above demonstration, the only estimator that ensures the between cluster variation is appropriately modeled and allows for the inclusion of higher level predictors is the CRE model where the cluster mean for the Level 1 variable is included; it may also surprise readers to see how well OLS, with cluster means, estimates such models too as long as a cluster-robust standard errors are used (thus researchers without multilevel modeling software can still estimate such multilevel models with basic software). Note too that in this demonstration, the unobserved effect, in terms of the ICC1 (Bliese, 1998), was very strong; however, results are substantively similar, though the bias is less pronounced, for smaller values of the ICC1 (see Appendix D).

Monte Carlo simulation

To demonstrate the effects of ignoring violations of the random effects assumption, we conduct a Monte Carlo simulation using two Level 1 predictors along with two Level 2 predictors. To focus on comparing the main modeling approaches currently in use based on our review with those that we think are best for model recovery, we chose ten modeling/estimation approaches for comparison: (1) the FE approach, estimated with GLS. (2-5), four RE approaches with either cluster (center CM) or grand-mean centering (center GM) both estimated with ML and GLS, and (6-10) five CRE approaches: The basic CRE model (Equation 7a) estimated with OLS, and the same model with and without cluster-mean centering (Equation 7a, 7b), both estimated with ML and GLS.

Simulation setup

Given our focus on endogeneity involving u_j, we generate data where u_j correlates with Level 1 and Level 2 regressors using the following model:

$$y_{ij} = \gamma_0 + \gamma_1 x1_{ij} + \gamma_2 x2_{ij} + \gamma_3 z1_j + \gamma_4 z2_j + u_j + e_{ij} \qquad \textit{Eq. 11}$$

γ_1 and γ_3 were varied as experimental conditions receiving values between -2 to 2 at increments of 1 and γ_2 and γ_4 were 2 in the base scenario. Both u_j and e_{ij} are normally distributed with mean = 0 and *SD* = 10 and 8 respectively. $x1_{ij}$ and $x2_{ij}$ were set to be correlated at .3 and the $z1_j$ and $z2_j$ variables were uncorrelated; for simplicity there were also no correlations between the levels.

The sample size was set based on the results of our systematic literature review reported later in the article. The Level 2 sample sizes were 20, 30, 50, 100, 500, and 1000 and the Level 1 sample sizes were 2, 5, 10, 20, and 30 producing a minimum of 40 and a maximum of 30,000 Level 1 observations. To model varying degree of endogeneity with the unobserved cluster effect, the correlation between u_j and the Level 1 variables was either .10, .30, or .50, corresponding to Cohen's (1992) classification of small, medium, and large correlations. The correlation between u_j and the Level 2 variables was 0, .20, or .40. These values were obtained by subtracting .1 from the previous correlations to generate a fully exogenous correlation; if consistent, all estimators should therefore recover the correct Level 2 estimates under the condition of strict exogeneity (having some endogeneity too in the Level 2 covariates is important to validate that all estimators will be biased and inconsistent in those conditions). To increase the generalizability of our results to conditions where the clustering effect is weak, we added an additional experimental factor where we increased the variance of the Level 1 variables ($x1_{ij}$, $x2_{ij}$, and e_{ij}) by either 1, 2, and 4. After this manipulation, across the models of strict exogeneity, the ICC1's of the dependent variable ranged from 0.085 to 0.885 in the population. These ranges amply cover, and also go beyond, what Bliese (2000, p. 361) considers in his experience to be normal ICC1's in field data.

The experiment was a $5 \times 5 \times 6 \times 5 \times 3 \times 3 \times 3$ full factorial with a total of 20,250 conditions. Each cell contained 1,000 replications. The simulation was done using R and the plm (Croissant

& Millo, 2008) and lme4 (Bates, et al., 2015) packages on a computer cluster. The R code is provided in Appendix E.

Results

We start the results section with the results for the Level 1 coefficients. Because the results for γ_1 and γ_2 were qualitatively the same, we only present the result for γ_1 (i.e. effect of $x1_{ij}$) that was varied as an experimental condition. Figure 1 shows the marginal effect of both Level 1 and Level 2 of sample size for the ten chosen analysis approaches over all endogeneity conditions. The mean estimation errors for the FE approach, all CRE approaches, and the RE approaches with cluster-mean centering are flat lines at zero indicating that these estimation approaches are unbiased regardless of Level 2 endogeneity and the unbiasedness does not depend on the sample size. The estimates using grand-mean centering are severely biased. These results thus demonstrate that RE model with grand-mean centering should never be used when the random effects assumption does not hold.

[insert Figures 1 & 2 here]

Figure 2 shows the mean squared estimation error (MSE), which for unbiased estimators is also the variance of the estimates. The results are again very similar as before; RE models with grand-mean centering stand out as the worst and the performance of the other four estimators is identical. The lines that converge toward zero when either of the sample sizes increases demonstrates the consistency property of these estimators. That the performance of the CRE and FE estimators are identical is not surprising given that GLS FE and the CRE have been proven to produce the same within estimate in linear models. Interestingly, out of the poorly performing RE approaches, the GLS RE estimator performs much more poorly than the ML estimator. However,

this is just a methodological curiosity, given that neither of these approaches should be applied when the random effects assumption fails.

The analysis of the effects of Level 2 variables reveal clear differences between the estimators. Again, because the results for γ_3 and γ_4 were qualitatively the same, we only present the result for γ_3 that was varied as an experimental condition (i.e. effect of $z1_j$). Because the FE approach cannot estimate the effects of Level 2 variables, we only focus on the RE and CRE approaches. We start again by inspecting the marginal effect of sample size on both levels over all endogeneity conditions on the bias of the estimators. As shown in Figure 3, all modeling approaches are severely biased. Clearly, none of the estimators are valid approaches for estimating the effects of endogenous Level 2 variables; to address these scenarios instrumental variables would be required as we mentioned earlier. This result underscores the importance of ensuring that modeled regressors must be exogenous.

[insert Figure 3 here]

Because the estimators were essentially useless when the Level 2 variables were endogenous with respect to the unobserved cluster effect u_j, we will now focus on the scenarios where the Level 2 variables are strictly exogenous, (i.e., uncorrelated with u_j and e_{ij}). We still retain the three other simulation factors, including the Level 1 endogeneity conditions and the sample sizes on both levels. Figure 4 shows that under these conditions, all nine estimators are essentially unbiased when the number of clusters of 50 or more, but the RE approaches have more difficulties when the number of clusters is smaller. However, even then the bias is fairly small.

Clear differences emerge in Figure 5 showing the efficiency of the estimators by their mean squared estimation errors. All CRE approaches outperform the RE approaches with cluster-

mean centering on the $x1_{ij}$ and $x2_{ij}$ variables by a clear margin. All estimators improve with the increasing number of clusters; however, the CRE estimators are more efficient in small samples. Moreover, increasing cluster size produces a clear improvement of the CRE approaches; the RE approaches are immune to this improvement and only react to increasing the number of clusters and not the cluster size (at least not at the Level 1 samples sizes we set), which could be quite impractical in some research contexts.

[insert Figures 4 and 5 here]

The efficiency differences can be explained by considering how the estimators use between cluster information of the Level 1 variables. In the CRE models, the cluster means contribute to explaining the variance of the dependent variable and increase in the explained variance leads to smaller variance of the estimator compared to the RE estimators that do not use cluster means. The small efficiency differences between the cluster and grand-mean centering approaches can be explained by the fact that cluster-mean centering further discards between cluster variation that would have been useful for explaining the dependent variable.

The simulation results provide two key takeaways:

1. To consistently estimate Level 1 effects, one should use one of the following estimators: (a) the FE estimator, the CRE approach implemented with either (b) OLS, (c) the RE (either GLS or ML) estimator with cluster means, or (d) the RE estimator with cluster-mean centering. The RE approach with grand-mean centering is clearly biased and inconsistent.
2. For models including Level 2 effects, to consistently estimate these, one should use either: (a) The OLS estimator with cluster means, or (b) the GLS or ML RE with cluster means (i.e. CRE). The RE approach with cluster-mean centering or grand-

mean centering have more absolute bias, particularly at smaller samples, and have higher variance in estimates regardless of whether the GLS or ML estimator is applied.

The above results have important implications, given our findings from the systematic literature review we conducted, which we present next. Of note too is the performance of OLS, which may come to a surprise to applied researchers and even methodologists, who have claimed, for instance that OLS regression should not be used to analyze panel data sets (Certo & Semadeni, 2006; Hofmann, 1997; Hofmann & Gavin, 1998). OLS can be used even if the random effects assumption holds, provided cluster means for the Level 1 variables are modeled and that the variance is estimated using a cluster-robust (sandwich) estimate.

Systematic review of the literature

At this point we have identified a potential problem with respect to how to estimate models having unobserved cluster effects, suggested how to fix the problem, and shown how this fix—as tested from rather extensive Monte Carlo simulations—works. The questions we now ask are: How do researchers estimate multilevel models? Do they consider the random effects assumption? Ultimately, how big is the endogeneity problem in the empirical literature? To address these questions, we conducted a systematic review of the empirical literature in top academic journals.

Review protocol

We first searched for journals listed in the categories "Psychology, Applied" and "Management" in Web of Science and ranked the top 10, based on their 5-year impact factor (as per April 2018). We selected journals publishing empirical studies that clearly relate to the management and applied psychology fields. Based on this screening procedure, we retained seven journals publishing macro-oriented research, micro-oriented research, or a mix of the two: *Academy of*

Management Journal, *Journal of Applied Psychology*, *Journal of Management*, *Journal of Organizational Behavior*, *Journal of Operations Management*, *The Leadership Quarterly*, and *Personnel Psychology*. To focus on the current practice, we considered every article published in 2016 and 2017, and included studies if they (a) modeled unobserved heterogeneity in some way (though we excluded categorical unobserved components or cross-lagged models which do not have an observation specific intercept); (b) included more than ten Level-2 clusters; and (3) had Level-1 varying predictors. Our initial pool contained 270 articles. To keep the coding workload reasonable we coded a sample of 150 articles. To do so, we randomly selected a proportional number of articles per journal except for journals that published fewer than 15 multilevel articles in which case we included all their articles (i.e., from *Journal of Operations Management, The Leadership Quarterly*, and *Personnel Psychology*). Because our initial sample contained a majority of micro-oriented articles (N=102), following a reviewer recommendation, we included two more macro journals: *Strategic Management Journal* and *Organization Science*. We randomly selected 54 articles from these two Journals to have a perfectly balanced sample of micro- and macro-oriented articles (final N=204).

We coded articles on descriptive aspects, such as the type of data, sample size at different levels, and the statistical program used; we also coded for certain qualitative aspects, such as the exogeneity of predictors, whether the random effects assumption was made (and if so, respected) and if the authors made causal claims. The latter point is important to note because our critique of current practice concerns researchers making causal claims (having policy implications) and not merely reporting on statistical associations.

The coding was done by the second author, and refined as required through constant discussion with the other authors as issues arose that were not foreseen in the coding manual (which was concomitantly refined and reapplied). To ensure the reliability of the coding, we

selected a random sample of 10 articles to be coded by a senior doctoral student with an excellent knowledge of econometrics. The 10 manuscripts generated 240 coding events. Expected agreement due to chance would have been 16.77%; however the coders agreed on 83.33% (i.e., 200 out of 240 coding events) and the agreement statistics $\kappa = .81$, SE = .03, $z = 30.57$, $p < .001$ indicate an agreement rate significantly better than chance and generally qualified as "substantial" (Landis & Koch, 1977). A summary of the coding manual is included in Appendix F. The coded data is available on the journal's website and linked to our article.

Results of systematic review

We start with descriptive statistics, and we refer interested readers to Appendixes G and H for more detailed summaries. Our sample of coded articles contains slightly more longitudinal data (49.51%) than hierarchical data (39.22%), with some articles having both types of data (11.27%). Sample sizes tend to fluctuate widely across studies, journals or type of dataset. The majority of articles used the RE approach with the ML estimator. We observed too that a large number of studies (43.14%) did not report the statistical program used for estimation. This issue is problematic because different software and pre-programmed commands use different defaults, potentially masking important information for readers and reviewers unaware of the specific settings used. For example, the Multilevel SEM technique in Mplus includes a latent variable for the cluster mean by default if (and only if) authors only include Level 1 regressors and do not specify any Level 2 variables.

By far, the most important issue pertains to the consistency of estimation. Thus, we reported which estimator was used and investigated how authors using an RE model ensured that its assumptions held. As we mentioned before researchers can (a) ensure the random effects assumption holds via a relevant statistical test, (b) model cluster means (CRE) or cluster-mean center their level-1 variables, or (c) use a FE estimator. Figure 6 displays the relative proportion

of articles per journal that followed each respective modeling strategy (note, we pooled estimators using cluster means (CRE), cluster centering and the different FE estimators into the category "FE and CRE" because these approaches do not make the random effects assumption). Overall, only 106 articles (60 macro and 46 micro) out of the total sample of 204 (i.e., 51.96%) either ensured that the random effects assumption held or applied the FE or CRE approach that do not require this assumption. In other words, almost half of the articles applied the RE approach without justifying its assumptions.

[insert Figure 6 here]

Additionally, we used our judgments to determine whether modeled predictors respected the "full exogeneity" criterion, that is, whether Level 1 and Level 2 predictors likely correlated with their respective error terms e_{ij} and u_j. Modeled predictors must be exogenous (i.e., the variable is manipulated, fixed, cyclical, or varies randomly in nature; see Antonakis et al., 2010 for examples). Unfortunately, the majority of articles used, what appeared to us to be endogenous Level 1 (79.90% of articles) and Level 2 (77.08%, i.e., 74 from 96 relevant articles that modeled Level 2 variables) regressors.

Table 6 shows an alarming result: Combining the exogeneity and the random effects assumption criteria showed that only 25 articles (12.25%) reported consistent estimates; most of the articles (n = 18) were from macro-oriented research. Such a large amount of potentially unreliable results indicates that the empirical literature still lags behind the recent methodological literature (Antonakis, et al., 2010; Halaby, 2004; Petersen, 2009). Yet, most articles (n = 137) made clear causal claims; however, only n = 22 of those articles did so on sound causal foundations. Thus, apart from making appropriate policy recommendations, researchers should pay attention to our suggestions also because articles with endogeneity threats are less well cited

than are those having a more causally defensible design (Antonakis, Bastardoz, Liu, & Schriesheim, 2014).

[insert Table 6 here]

Discussion

Our critical analysis of the various multilevel modeling methods highlights a key finding: There are proven ways to accurately estimate multilevel models without jeopardizing estimating consistency. We focused on relatively simple models having Level 1 and Level 2 effects, but the issues we highlight are general and concern more complex models too. Thus, our findings and recommendations have far-reaching conclusions that apply to both micro and macro-oriented researchers. Interestingly, macro-oriented researchers are doing a bit better than micro-oriented researchers; the former are closer to economics both in method and substance, which probably explains the differences we see. Of course, precision in measurement is also important, which is more the providence of micro research. Both these issues, endogeneity and measurement, are important, but seemingly not sufficiently covered in doctoral training (Aiken, West, & Millsap, 2008; Antonakis, et al., 2010).

Violations of the random effects assumption

Regrettably, the results of our systematic review show that most researchers are unaware of the assumptions required for estimating multilevel models. This observation is particularly troublesome given the fact that the random effects assumption is testable and can be relaxed by applying the CRE model. More needs to be done to make researchers aware of the key issues; it *is critical that researchers are appropriately trained to understand the conditions required to consistently and unbiasedly estimate multilevel models*. However, it seems from our anecdotal observations that oftentimes doctoral courses focus on the "how to" of using specialized programs (e.g., HLM) instead of giving students the needed mathematical undergirding to

understand what endogeneity is and how it must be dealt with appropriately in multilevel models and beyond. In the reviewed articles, researchers often note that they used a multilevel procedure so as to not violate the independence assumption of observations. This assumption is easily handled by using a cluster-robust standard errors, yet researchers appear to be unware of even such remedial matters (McNeish, et al., 2017); of course, a far greater problem is first obtaining consistent and unbiased estimates by correctly modeling the structure of the data.

Finally, there is an argument that the use of an RE approach over FE (or CRE) is sometimes preferable when the Hausman test fails (Clark & Linzer, 2015). This argument is based on the observation that if the random effects assumption is violated only trivially, the bias of the RE approach will also be trivial and RE should be favored over FE given the efficiency gains of the former. The reasoning here is that when the sample size is large, the Hausman test (and any test for that matter) detects trivially small effects. However, in such large sample, the difference between the RE and FE estimates is trivial too given that both are consistent. Moreover, in large samples efficiency is less of a concern because all consistent estimators are precise enough. Thus, as a standard practice, we recommend to never use an RE approach when a test indicates that its assumption fails.

Confusion on centering

Our review showed that centering decisions are not correctly taken by researchers. We fear that a key reason may be because methodological writings on the topic have presented misleading guidelines. For instance, recommendations about centering have typically revolved around stylistic issues regarding the interpretation of a parameter as a function of the measurement properties of the regressor. Surprisingly too, the literature has suggested that there is no statistical test to guide the type of centering approach—and thus implicitly which modeling procedure —to use. Consider for instance the following:

1. For Raudenbush and Bryk (2002), "if an X_{ij} value of zero is not meaningful, then the researcher may want to transform X_{ij}, or choose a location for X_{ij} that will render [its coefficient] more meaningful" (p. 32).
2. According to Kreft et al. (1995), both centering options [grand or cluster] are "statistically sound ways to improve parameter estimation [and] the choice between the two options for centering can only be made on a theoretical basis" (p. 1); that centering "may facilitate interpretation" (p. 2); and that grand-mean centering provides "computational advantages" by reducing "multicollinearity" (p. 10).
3. Per Hofmann (1997) centering renders "intercepts more interpretable" and that researchers must "consider their overarching theoretical paradigm and from that discern what centering option best represents their paradigm" (p. 738).
4. Hofmann and Gavin (1998) noted that because scales usually do not have natural zero points, centering should be done to "render the intercept term more interpretable or meaningful" (p. 626); moreover, all "centering options are statistically appropriate [and that] the choice of centering options must be a function of the conceptual paradigm and research question under investigation" (p. 638); and finally, grand-mean centering reduces "potential problems associated with multicollinearity" (p. 638).
5. Enders and Tofighi (2007) stated "both [types of centering] are appropriate in certain circumstances and are inappropriate in others" (p. 127); more importantly, cluster-mean centering "may be the most appropriate form of centering in situations in which the primary substantive interest involves a Level 1 . . . predictor," and that grand-mean centering "is the method of choice for assessing the impact of cluster-level variables, controlling for Level 1 covariate" (p. 128)

Whereas the above suggestions have been nuanced (e.g., Raudenbush & Bryk, 2002, see pp. 139 & 183), these recommendations with respect to centering decisions appear to have mislead researchers to think that centering depends on how the regressors are scaled. These recommendations are incorrect for two reasons. First, grand-mean centering (in linear models) will only change the estimated intercept; thus, applying this type of centering, as many did (10.78% of article, most of which were micro articles) is futile. Not only is grand-mean centering useless (the intercept itself is rarely of interest), it can also be harmful when researchers interested in calculating and plotting marginal predictions for different combinations of the explanatory variables. If these predictions are calculated using the predictors on their original metric from a model estimated with grand-mean centering, the predictions will be systematically incorrect (Dawson, 2014).

Second, the use of cluster-mean centering, which is more popular (19.61% of the articles used it), has some merit, but this technique has nothing to do with the scale of the predictors. Cluster-mean centering produces a within effect that is free of the endogeneity problem discussed in this article. Indeed, cluster-mean centering is the first step in the GSL FE estimator that is typically recommended as the default option in econometrics (McNeish & Kelley, 2018). However, given that the CRE approach can accomplish the same as cluster-mean centering and provides more flexibility in modeling as well as providing the contextual effect, we see the CRE approach as a much more attractive alternative to cluster-mean centering. Nevertheless, there is one scenario where cluster-mean centering is required: If a researcher is interested in estimating the between effect instead of the contextual effect, centering must be applied to the Level 1 variables *along with including the cluster means of Level 1 variables*. However, our review showed that researchers rarely included or reported on having used cluster means as control variables (i.e., over 90% of articles).

Extending the CRE approach to random slope models and three level models

Whereas our review of the literature and demonstration focuses on the unobserved cluster-specific intercept, the CRE approach we advocate and the tests for random effects assumption extends readily to other random effect models as well (Bell & Jones, 2015). Consider the CRE approach in the context of Eq. 1 presented in the multilevel format:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij} \quad \} Level\ 1\ equation \qquad Eq.\ 12a$$

$$\left.\begin{array}{l} \beta_{0j} = \gamma_{00} + \gamma_{01}\bar{x}_j + u_{0j} \\ \beta_{1j} = \gamma_{10} + \gamma_{11}\bar{x}_j + u_{1j} \end{array}\right\} Level\ 2\ equations \qquad \begin{array}{l} Eq.\ 12b \\ Eq.\ 12c \end{array}$$

The CRE approach is implemented by including cluster means of the Level 1 variables to all Level 2 equations. In the mixed format the same can be expressed as:

$$y_{ij} = \underbrace{\gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}\bar{x}_j + \gamma_{11}\bar{x}_j x_{ij}}_{fixed\ part} + \underbrace{x_{ij}u_{1j} + u_{0j} + e_{ij}}_{random\ part} \qquad Eq.\ 13$$

The equation shows that when the random effect of interests affects a regression slope instead of an intercept, the CRE approach can be implemented by adding an interaction between the original variable and its cluster mean as a control. It is important to note that even though it is a common practice to center variables before calculating interactions, this procedure should not be done in this particular case. More generally, centering when forming interactions is largely misunderstood and a inutile practice (Dalal & Zickar, 2012; Kromrey & Foster-Johnson, 1998).

Finally, beyond more complex Level 2 models, the CRE approach can be readily extended to three level models and beyond. In the case of a three-level model, the equations for a random intercept model would be:

$$y_{ijk} = \beta_{0jk} + \beta_1 x_{ijk} + e_{ijk} \quad \} Level\ 1\ equation \qquad Eq.\ 14a$$

$$\beta_{0jk} = \gamma_{00k} + \gamma_{01k}\bar{x}_{jk} + u_{0jk} \} Level\ 2\ equation \qquad Eq.\ 14b$$

$$\beta_{00k} = \gamma_{000} + \gamma_{001}\bar{x}_k + u_{00k} \; \} Level\ 3\ equation \qquad Eq.\ 14c$$

where *k* indicates the third level of clustering. In the mixed effects format, this translates to a model with two cluster means of *x*, calculated for both Level 2 and Level 3 clusters. However, considering that the sample sizes on Level 3 and higher tend to be small, estimating the contextual effects on these levels would be imprecise and also often of not direct interest and therefore using an FE approach by including dummies to indicate the Level 3 units presents a compelling alternative (cf. McNeish & Wentzel, 2017).

Recommendations and Conclusions

Following our critical analysis, simulation demonstrations, and review, we summarize key points that applied researchers should adopt to correctly estimate multilevel models. The recommendations are summarized in Figure 7 as a decision chart. More specifically:

1. Use the CRE approach as the default estimator particularly if effects of higher-level variables (e.g., Level 2 or higher) are of interest; the estimator used, OLS (with cluster robust estimation of standard errors), ML, or GLS is immaterial. Although the FE estimate is consistent for the within effect, and even though the CRE approach produces the same within-estimates, ideally use CRE as the default given the extra information it provides in estimating the contextual effect.
2. Centering Level 1 data is unnecessary in most cases and complicates calculating marginal predictions and marginal effects. It should thus be avoided unless there is a good reason to center. More particularly:
    - Grand-mean centering is completely useless; as such, it should always be avoided.
    - Cluster-mean centering is required along with the use of cluster means if the interest is in estimating the between effect instead of the contextual effect along with the within effect.

- Never cluster-mean center the dependent variable outside the GLS procedure because this will bias the standard errors (Wooldridge, 2002 pp. 269-272).

3. Use the RE approach only if the random effects assumption is empirically assessed and thus justified. Researchers can test the random effects assumption with either a Hausman, likelihood ratio, *F* or Wald test (or any other test that is valid for this purpose).

Of course that endogeneity with respect to the cluster effect has been addressed with CRE does not mean that all endogeneity issues have been eliminated. Researchers must consider what model they are estimating, and ensure that the modeled independent variables are exogenous. Moreover, one final point to bear in mind, which we did not cover, is that when estimating panel models, researchers should examine whether clustered-robust standard errors change inference, because these are also robust for autocorrelation (Angrist & Pischke, 2008, see Ch. 9); if substantially different these standard errors should be reported instead of the conventional ones (Cameron, Gelbach, & Miller, 2011; Cameron & Miller, 2015).

[insert Figure 7 here]

To conclude, although our findings are not particularly encouraging, researchers and educational institutions must undertake concerted efforts to redress the situation. We are optimistic that with time, research practice will improve and better inform policy. Other disciplines have gone through similar growing pains, like economics did when it faced its endogeneity demons. The casual identification revolution shook it to the core, but economics emerged stronger as a discipline, with a unified methodological paradigm, and better research practice (e.g., see Angrist & Pischke, 2010). We can too.

References

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology - Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63*(1), 32-50.

Allison, P. D. (2009). *Fixed effects regression models*. Los Angeles: Sage publications.

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton: Princeton University Press.

Angrist, J. D., & Pischke, J.-S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics. *Journal of Economic Perspectives, 24*(2), 3-30.

Angrist, J. D., & Pischke, J.-S. (2014). *Mastering metrics: The path from cause to effect*. Princeton: Princeton University Press.

Antonakis, J. (2011). Predictors of leadership: The usual suspects and the suspect traits. In A. Bryman, D. Collinson, K. Grint, B. Jackson & M. Uhl-Bien (Eds.), *Sage Handbook of Leadership* (pp. 269-285). Thousand Oaks: Sage Publications.

Antonakis, J., Bastardoz, N., Liu, Y., & Schriesheim, C. A. (2014). What makes articles highly cited? *The Leadership Quarterly, 25*(1), 152-179.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly, 21*, 1086-1120.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2014). Causality and endogeneity: Problems and solutions. In D. V. Day (Ed.), *The Oxford Handbook of Leadership and Organizations* (pp. 93-117). New York: Oxford University Press.

Arellano, M. (1993). On the testing of correlated effects with panel data. *Journal of Econometrics, 59*(1-2), 87-97.

Ballinger, G. A. (2004). Using generalized estimating equations for longitudinal data analysis. *Organizational Research Methods, 7*(2), 127-150.

Bascle, G. (2008). Controlling for endogeneity with instrumental variables in strategic management research. *Strategic Organization, 6*(3), 285-327.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Bell, A., & Jones, K. (2015). Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data. *Political Science Research and Methods, 3*(1), 133-153.

Bliese, P. D. (1998). Group Size, ICC Values, and Group-Level Correlations: A Simulation. *Organizational Research Methods, 1*(4), 355-373.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for aggregation and analysis. In S. W. J. Kozlowski & K. J. Klein (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349-381). San Francisco, CA: Jossey-Bass.

Bliese, P. D., & Ployhart, R. E. (2002). Growth modeling using random coefficient models: Model building, testing, and illustrations. *Organizational Research Methods, 5*(4), 362-387.

Bollen, K. A. (2012). Instrumental Variables in Sociology and the Social Sciences. *Annual Review of Sociology, 38*(1), 37-72.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust Inference With Multiway Clustering. *Journal of Business & Economic Statistics, 29*(2), 238-249.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources, 50*(2), 317-372.

Certo, S. T., & Semadeni, M. (2006). Strategy Research and Panel Data: Evidence and Implications. *Journal of Management, 32*(3), 449-471.

Certo, S. T., Withers, M. C., & Semadeni, M. (2017). A tale of two effects: Using longitudinal data to compare within‐and between‐firm effects. *Strategic Management Journal, 38*(7), 1536-1556.

Clark, T. S., & Linzer, D. A. (2015). Should I Use Fixed or Random Effects? *Political Science Research and Methods, 3*(2), 399-408.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Croissant, Y., & Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software, 27*(2), 1-43.

Curran, P. J., & Bauer, D. J. (2011). The disaggregation of within-person and between-person effects in longitudinal models of change. *Annual Review of Psychology, 62*, 583-619.

Dalal, D. K., & Zickar, M. J. (2012). Some Common Myths About Centering Predictor Variables in Moderated Multiple Regression and Polynomial Regression. *Organizational Research Methods, 15*(3), 339-362.

Dawson, J. F. (2014). Moderation in Management Research: What, Why, When, and How. *Journal of Business and Psychology, 29*(1), 1-19.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121-138.

Gennetian, L. A., Magnuson, K., & Morris, P. A. (2008). From statistical associations to causation: What developmentalists can learn from instrumental variables techniques coupled with experimental data. *Developmental Psychology, 44*(2), 381-394.

Greene, W. H. (2012). *Econometric analysis*. Boston: Prentice Hall.

Guo, G. (2017). Demystifying variance in performance: A longitudinal multilevel perspective. *Strategic Management Journal, 38*(6), 1327-1342.

Halaby, C. N. (2004). Panel models in sociological research: Theory into practice. *Annual Review of Sociology, 30*, 507-544.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica, 46*(6), 1251-1271.

Helson, R., Jones, C., & Kwan, V. S. Y. (2002). Personality change over 40 years of adulthood: Hierarchical linear modeling analyses of two longitudinal samples. *Journal of Personality and Social Psychology, 83*(3), 752-766.

Hofmann, D. A. (1997). An overview of the logic and rationale of hierarchical linear models. *Journal of Management, 23*(6), 723-744.

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management, 24*(5), 623-641.

Holcomb, T. R., Combs, J. G., Sirmon, D. G., & Sexton, J. (2010). Modeling levels and time in entrepreneurship research: An illustration with growth strategies and post-IPO performance. *Organizational Research Methods, 13*(2), 348-389.

Kreft, I. G., De Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30*(1), 1-21.

Kromrey, J. D., & Foster-Johnson, L. (1998). Mean centering in moderated multiple regression: Much ado about nothing. *Educational and Psychological Measurement, 58*(1), 42-67.

La Porta, R., Lopez-De-Silanes, F., & Shleifer, A. (2008). The economic consequences of legal origins. *Journal of Economic Literature, 46*(2), 285-332.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Larcker, D. F., & Rusticus, T. O. (2010). On the use of instrumental variables in accounting research. *Journal of Accounting and Economics, 49*(3), 186-205.

Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist, 35*(2), 125-141.

McNeish, D., & Kelley, K. (2018). Fixed effects models versus mixed effects models for clustered data: Reviewing the approaches, disentangling the differences, and making recommendations. *Psychological Methods, 24*(1), 20-35.

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods, 22*(1), 114-140.

McNeish, D., & Wentzel, K. R. (2017). Accommodating small sample sizes in three-level models when the third level is incidental. *Multivariate Behavioral Research, 52*(2), 200-215.

Mundlak, Y. (1978). Pooling of Time-Series and Cross-Section Data. *Econometrica, 46*(1), 69-85.

Neuhaus, J. M., & Kalbfleisch, J. D. (1998). Between- and Within-Cluster Covariate Effects in the Analysis of Clustered Data. *Biometrics, 54*(2), 638-645.

Petersen, M. A. (2009). Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches. *The Review of Financial Studies, 22*(1), 435-480.

Pustejovsky, J. E., & Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics, 36*(4), 672-683.

Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*. College Station, TX: Stata Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks: Sage.

Sachs, J. D. (2003). Institutions don't rule - direct effects of geography on per capita income. *Working Paper No. 9490, February. National Bureau of Economic Research*.

Schaffer, M. E., & Stillman, S. (2006). Xtoverid: Stata module to calculate tests of overidentifying restrictions after xtreg, xtivreg, xtivreg2 and xthtaylor. *http://ideas.repec.org/c/boc/bocode/s456779.html*.

Schunck, R. (2013). Within and between estimates in random-effects models: Advantages and drawbacks of correlated random effects and hybrid models. *The Stata Journal, 13*(1), 65-76.

StataCorp. (2017). *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LP.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (5th ed.). Mason, OH: South-Western Cengage Learning.

Yammarino, F. J., & Dansereau, F. (2011). Multi-level issues in evolutionary theory, organization science, and leadership. *The Leadership Quarterly, 22*, 1042-1057.

Table 1 Summary of modeling approaches and estimation techniques

Approach to contextual effects	Approach to unobserved heterogeneity	Specific techniques	Critical features			Estimator
			RE assumpt.	Effi-cient	L2 var.	
FE approaches						
Eliminated	Eliminated	Dummy variables	No	No	No	OLS ML
		Cluster-means centering *all* variables (GLS FE)	No	No	No	GLS FE
	Modeled	Cluster-mean centering IV but not DV and using random intercept	No	No	No	ML GLS RE
	Taken into account in estimation	Cluster-mean centering IV but not DV and using cluster robust SEs	No	No	No	OLS or GEE with cluster robust SEs
RE approaches						
Assumed to not exist	Modeled	Quasi-mean centering all variables ("GLS RE")	Yes	Yes	Yes	GLS RE
		Random intercept model	Yes	Yes	Yes	ML RE
	Taken into account in estimation	Clustered standard errors	Yes	No	Yes	OLS or GEE
CRE approaches						
Modeled	Modeled	Cluster means as controls and quasi-mean centering	No	No	Yes	GLS RE
		Cluster means as control and random intercept model	No	No	Yes	ML
	Taken into account in estimation	Cluster means as controls and clustered standard errors	No	No	Yes	OLS or GEE with cluster robust SEs

L2 var. = Level 2 variables; Estimation commands: OLS: regress (Stata); lm (R) ML: mixed (Stata); nlme; lmer (R) GLS RE: xtreg, re; xtreg, mle (Stata); plm(type ="random") (R) GLS FE: xtreg, fe (Stata); plm(model="within") (R); GEE: xtgee (Stata), gee (R); Cluster robust SE: vce(cluster) (Stata), vcovCR (R).

Table 2: Examples of within and contextual effects

	Within effect	Contextual effect
Vaccinations on health	Positive: Getting a vaccination decreases individuals risk of contracting a disease	Positive: Increasing the vaccination rate in the community where one lives decreases ones risk of contracting a disease (herd immunity)
Overfishing on profits	Positive: If a professional angler exceeds her fishing quota, her profits will increase because of larger catch	Negative: If there is overfishing on a lake, the profits of all anglers will decrease because of smaller catches
Innovativeness on competitive advantage	Positive: If a firm is innovative, it can develop valuable capabilities that lead to competitive advantage	Negative: If everyone in an industry innovates a lot, innovations are no longer rare and are less likely to lead to competitive advantage, but a non-innovating firm may have competitive disadvantage.
Gender on performance	Zero: One's gender has no impact on one's performance as a team member	Inverted U-shape: Teams gender composition affects the team performance; a team with half men and half women works best

Note: The between effect is the sum of these two effects and tells how the mean performance depends on the mean predictor across different contexts (e.g. teams, industries, etc.).

Table 3: Setting up the data for a correlated random effects (CRE) model and other approaches

N_j	n_{ij}	y_{ij}	z_j	x_{ij}	x_cl_j	$x - x\_cl_j$	$x - \bar{x}$	*d1*	*d2*	*d3*	*d4*	*d5*
1	1	31.09	1	3.88	4.09	-.21	1.36	0	0	0	0	0
1	2	34.16	1	4.31	4.09	.22	1.79	0	0	0	0	0
2	1	33.59	0	6.21	5.09	1.12	3.69	1	0	0	0	0
2	2	32.74	0	3.97	5.09	-1.12	1.45	1	0	0	0	0
3	1	26.83	0	-1.90	-1.08	-.82	-4.42	0	1	0	0	0
3	2	21.74	0	-.25	-1.08	.83	-2.77	0	1	0	0	0
4	1	19.61	1	-3.10	-3.37	.27	-5.62	0	0	1	0	0
4	2	16.07	1	-3.64	-3.37	-.27	-6.16	0	0	1	0	0
5	1	37.84	0	7.72	7.00	.72	5.20	0	0	0	1	0
5	2	35.28	0	6.29	7.00	-.71	3.77	0	0	0	1	0
6	1	32.66	1	3.96	3.36	.60	1.44	0	0	0	0	1
6	2	26.26	1	2.76	3.36	-.60	.24	0	0	0	0	1

Note: For the first cluster (N_j=1), the cluster mean for x_{ij} is (3.88 + 4.31)/2 = 4.09. The column $x - x_cl_j$we reports the group-mean centered data and the column $x - \bar{x}$ the grand-mean centered data (note, grand mean $\bar{x} = 2.52$, rounded). The variable z_j is a Level 2 variable. The remaining columns are the $k - 1$ dummy variables, where k = number of clusters leaving first cluster as a reference.

Table 4 Summary of tests of random effects assumption

Test of random effects assumption	Description	Decision rule	Critical features
Hausman	Compares an efficient (RE) to a consistent (FE or CRE) estimator	If estimated coefficients differ significantly, authors need to use the consistent estimator (i.e., FE or CRE) needs to be used otherwise the efficient estimator (i.e., RE) should be used.	Cannot be used with robust standard errors
Likelihood ratio-test	Compares two nested models: One model with cluster-means and the other model without cluster-means	If the likelihood ratio test is significant, cluster means have to be retained in the model.	Requires the use of ML estimation (i.e., cannot be used to compare a GLS FE with a GLS RE)
F- or Wald test	Tests the significance of contextual effects of cluster means	If the *F*- or Wald test is significant, cluster means have to be retained in the model.	Can be used with robust standard errors

Table 5: Comparison of different estimators and centering with data having a high ICC1

Variable	(1) OLS	(2) OLS	(3) GLS FE	(4) OLS	(5) OLS	(6) GLS RE	(7) ML RE	(8) ML RE	(9) GLS RE	(10) GLS RE	(11) ML RE
x	**2.97*****	.51***	.51***	.51***	.51***	.51***	.51***		**1.23*****		
	(29.37)	(17.53)	(18.48)	(18.48)	(18.48)	(18.48)	(17.19)		(23.24)		
z	**-.39*****				-.47***	-.47***	-.47***	-.47***	**-.33*****	**-.28****	**-.33*****
	(6.35)				(13.58)	(13.58)	(12.93)	(-12.93)	(3.39)	(2.30)	(3.39)
$\bar{x}_j$				4.52***	4.58***	4.58***	4.58***	5.09***			
				(51.19)	(59.28)	(59.28)	(60.76)	(73.56)			
$x - \bar{x}_j$								.51***		.51***	
								(17.19)		(18.48)	
$x - \bar{x}$											**1.23*****
											(23.24)
Dummies		Included									
Constant	-2.85*	-12.29***	-12.62***	-12.56***	-.89	-.89	-.89	-.89	-4.45*	-5.58*	-4.47*
	(1.85)	(1053.31)	(36061.00)	(153.85)	(1.04)	(1.04)	(.98)	(.98)	(1.85)	(1.82)	(1.85)

Note: Cluster robust *t*-statistics in parentheses (except for ML estimator); n = 5,000 observations (clustered under N = 500 leaders). OLS = Ordinary Least Squares estimator; GLS FE = GLS fixed effects estimator; GLS RE = GLS random effects estimator; ML RE = maximum likelihood estimator with RE model; the true coefficient for *x* is .50 and that of z is -.50 in the population. ***p < .01, **p < .05, *p < .10. Note, estimates that are substantially erroneous are emphasized (bolded-underlined).

Table 6: Coding of articles from journals with respect to consistency of estimated coefficients

	AMJ (N=44)	*SMJ (N=38)*	*JAP (N=27)*	*JOM (N=22)*	*JOB (N=19)*	*OS (N=16)*	*LQ (N=14)*	*PP (N=13)*	*JOpM (N=11)*	***All (N=204)***
L1 variable(s) exogenous	3 (7%)	5 (13%)	5 (19%)	0 (0%)	2 (11%)	4 (25%)	4 (29%)	0 (0%)	0 (0%)	**23 (11%)**
L1 variable(s) endogenous	35 (80%)	25 (66%)	22 (81%)	18 (82%)	17 (89%)	12 (75%)	10 (71%)	13 (100%)	11 (100%)	**163 (80%)**
L1 variable(s) instrumented	6 (14%)	8 (21%)	0 (0%)	4 (18%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	**18 (9%)**
L2 variable(s) exogenous	2 (5%)	1 (3%)	7 (26%)	2 (9%)	1 (5%)	2 (13%)	4 (29%)	2 (15%)	1 (9.%)	**22 (11%)**
L2 variable(s) endogenous	23 (52%)	3 (8%)	10 (37%)	9 (41%)	8 (42%)	3 (19%)	10 (71%)	6 (46%)	2 (18%)	**74 (36%)**
No L2 variable	19 (43%)	34 (89%)	10 (37%)	11 (50%)	10 (53%)	11 (69%)	0 (0%)	5 (38%)	8 (73%)	**108 (53%)**
RE assumption not made	17 (39%)	24 (63%)	14 (52%)	10 (45%)	9 (47%)	8 (50%)	5 (36%)	4 (31%)	7 (64%)	**98 (48%)**
RE empirically demonstrated to hold	3 (7%)	1 (3%)	0 (0%)	1 (5%)	0 (0%)	3 (19%)	0 (0%)	0 (0%)	0 (0%)	**8 (4%)**
RE assumption made and not respected	23 (52%)	13 (34%)	13 (48%)	11 (50%)	10 (53%)	4 (25%)	9 (64%)	9 (69%)	4 (36%)	**96 (47%)**
Unable to determine	1 (2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (6%)	0 (0%)	0 (0%)	0 (0%)	**2 (1%)**
Consistent estimates	**4 (9%)**	**10 (26%)**	**4 (15%)**	**2 (9%)**	**1 (5%)**	**2 (13%)**	**2 (14%)**	**0 (0%)**	**0 (0%)**	**25 (12%)**

Note: L1 = Level 1; L2 = Level 2; N=Number of articles included in our sample, broken down by journal. AMJ=Academy of Management Journal; SMJ=Strategic Management Journal; JAP=Journal of Applied Psychology; JOM=Journal of Management; JOB=Journal of Organizational Behavior; OS=Organization Science; LQ = Leadership Quarterly; PP=Personnel Psychology; JOpM=Journal of Operations Management; percentages are rounded.

Figure 1: Marginal effect of both levels of sample size on estimates of effect of *x1* over all endogeneity conditions

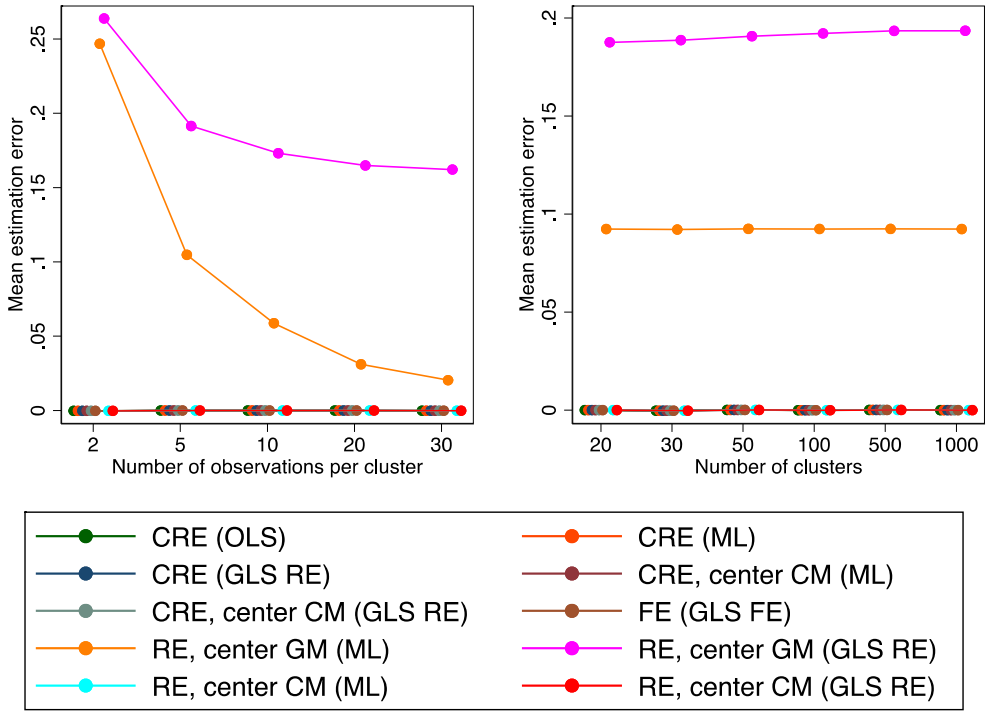


Figure 2: Marginal effect of both levels of sample size on squared estimation error of effect of *x1* over all endogeneity conditions

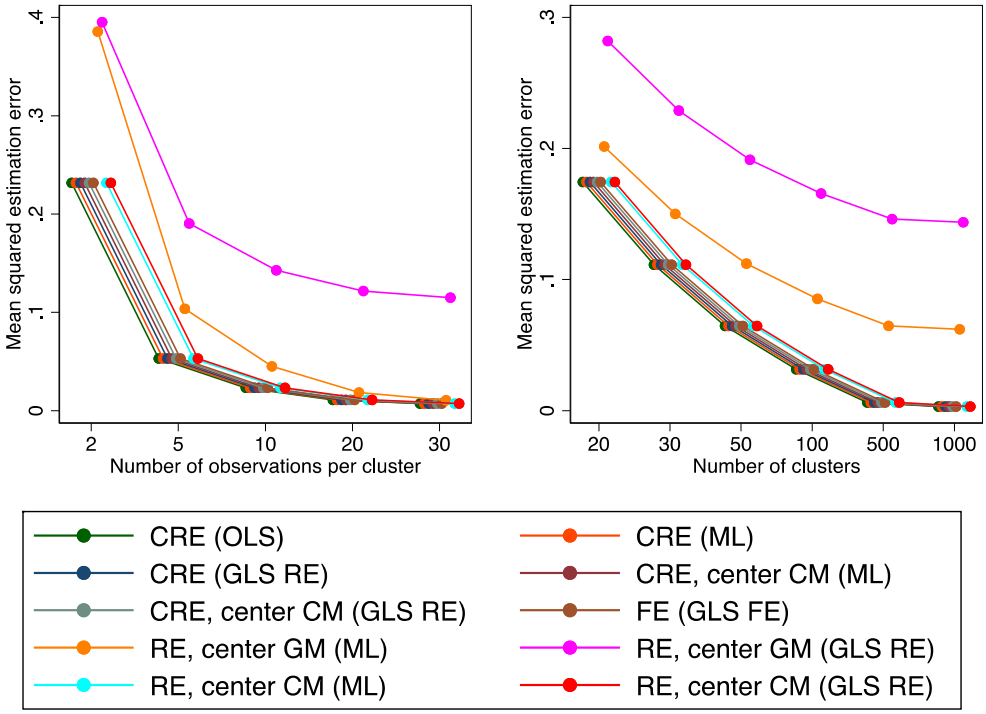


Figure 3: Marginal effect of both levels of sample size on estimates of effect of *z1* over all endogeneity conditions

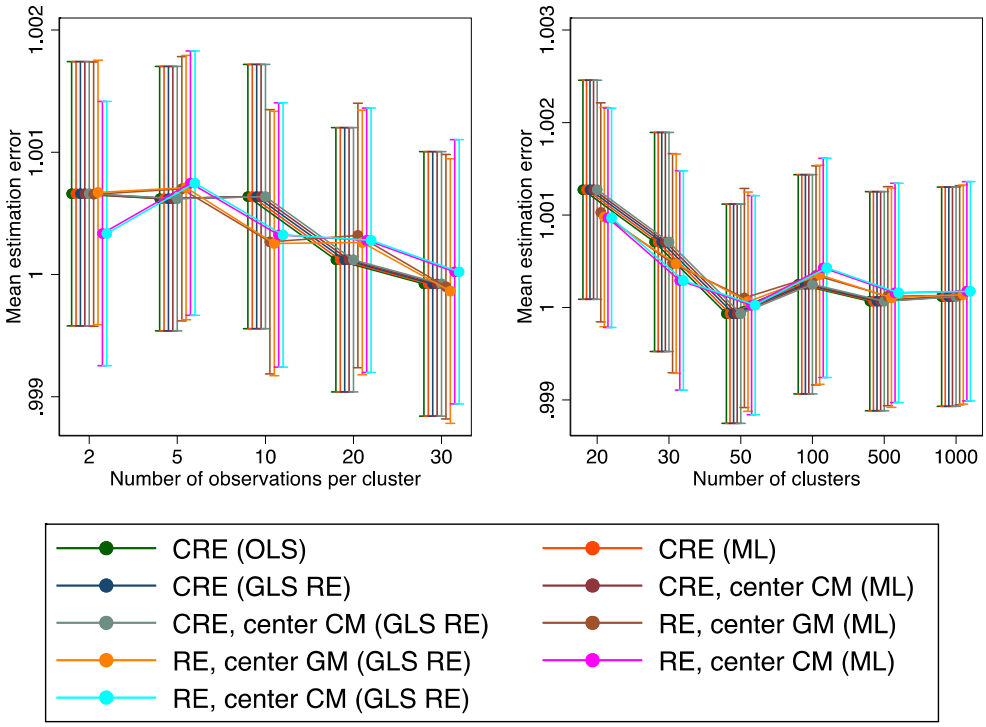


Figure 4: Marginal effect of both levels of sample size on estimates of effect of *z1* over all Level 1 endogeneity conditions when Level 2 endogeneity is zero

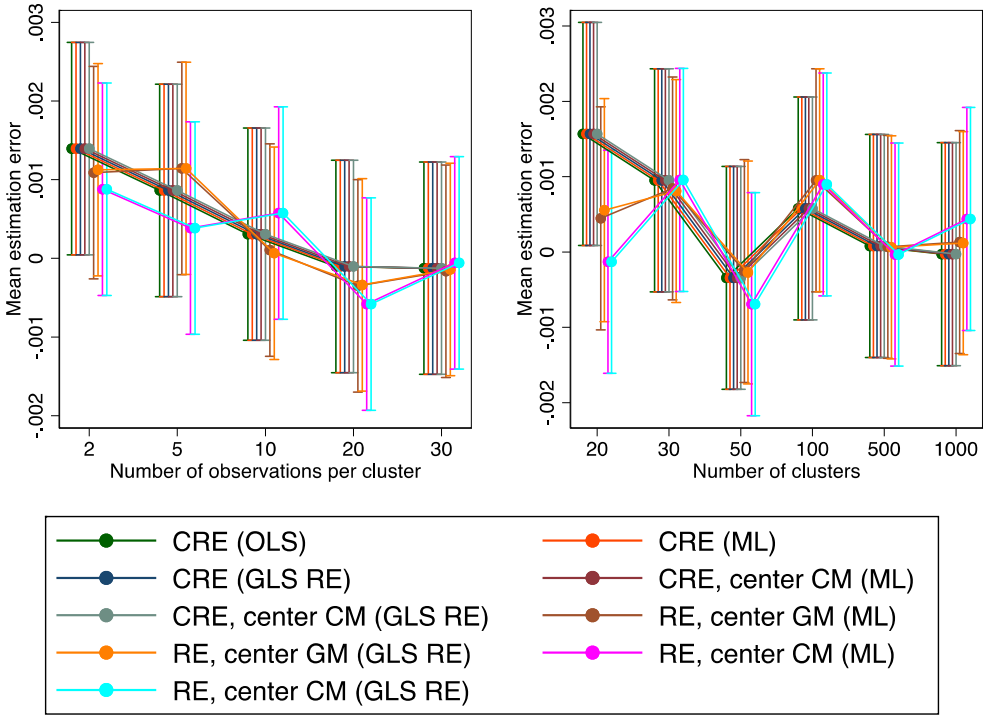


Figure 5: Marginal effect of both levels of sample size on squared estimation error of effect of *z1* over all Level 1 endogeneity conditions when Level 2 endogeneity is zero

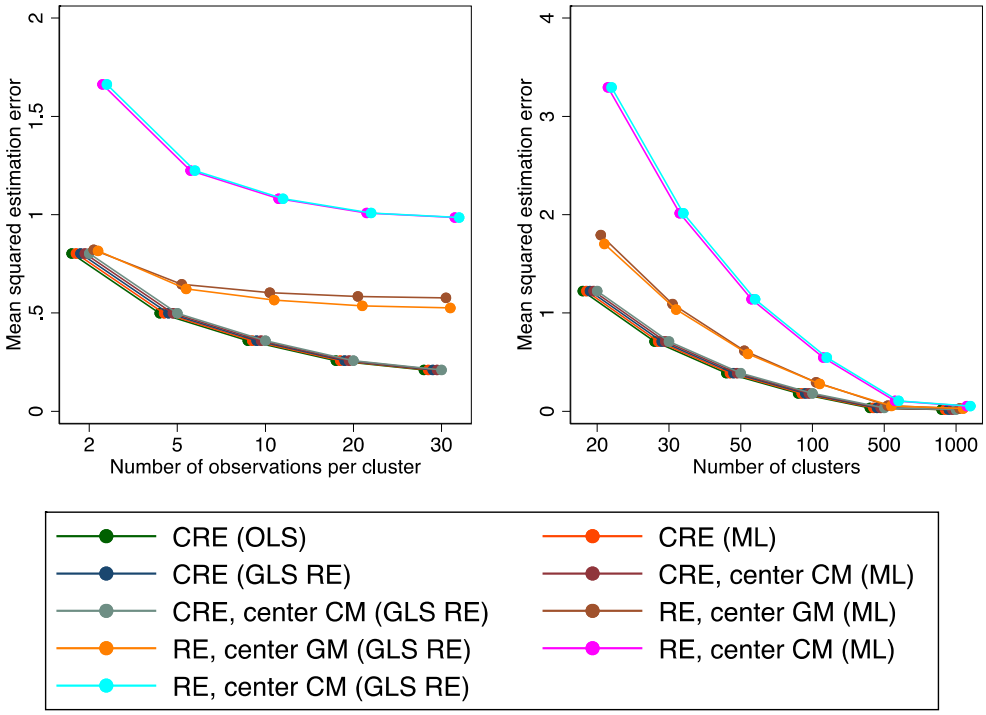
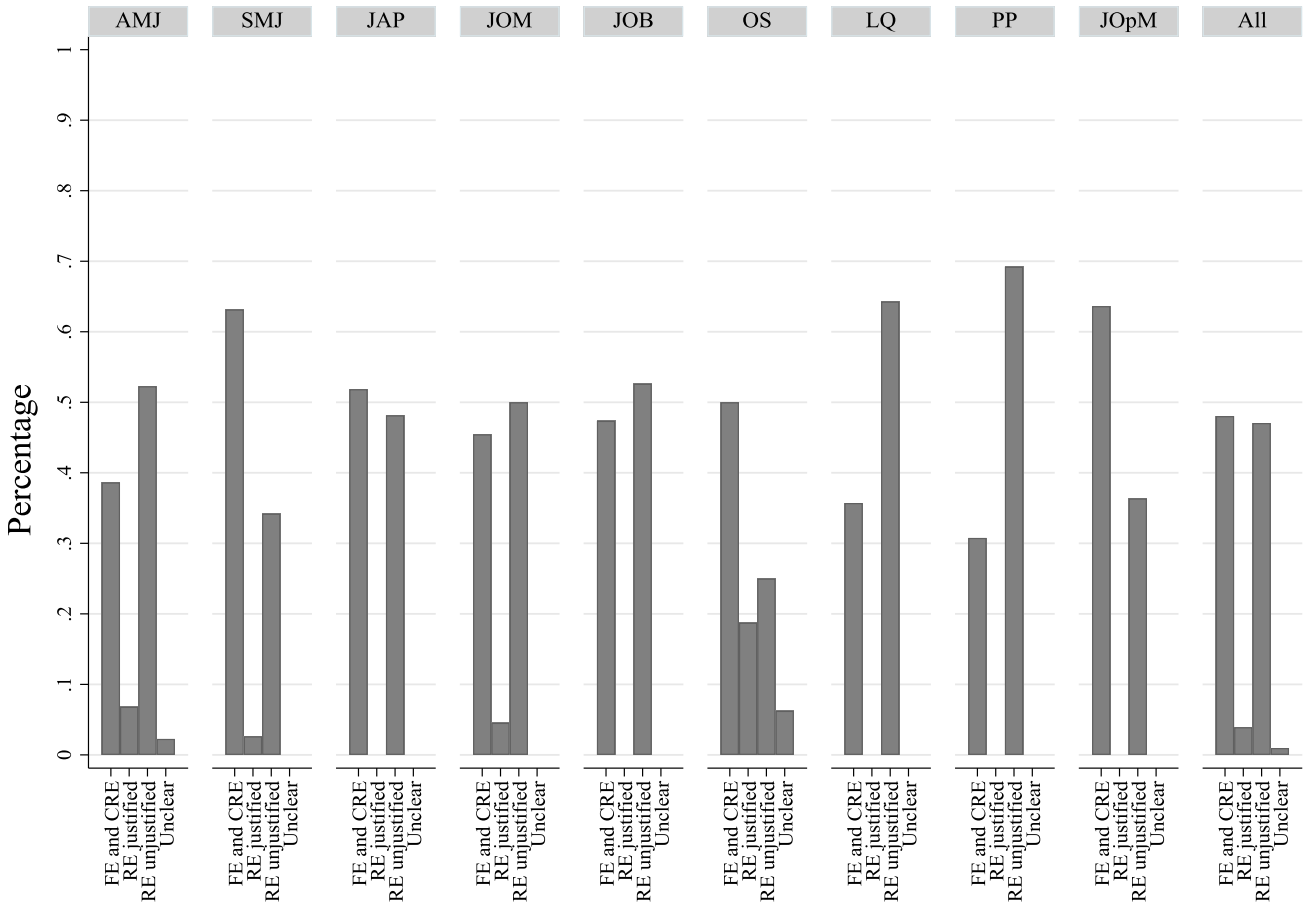
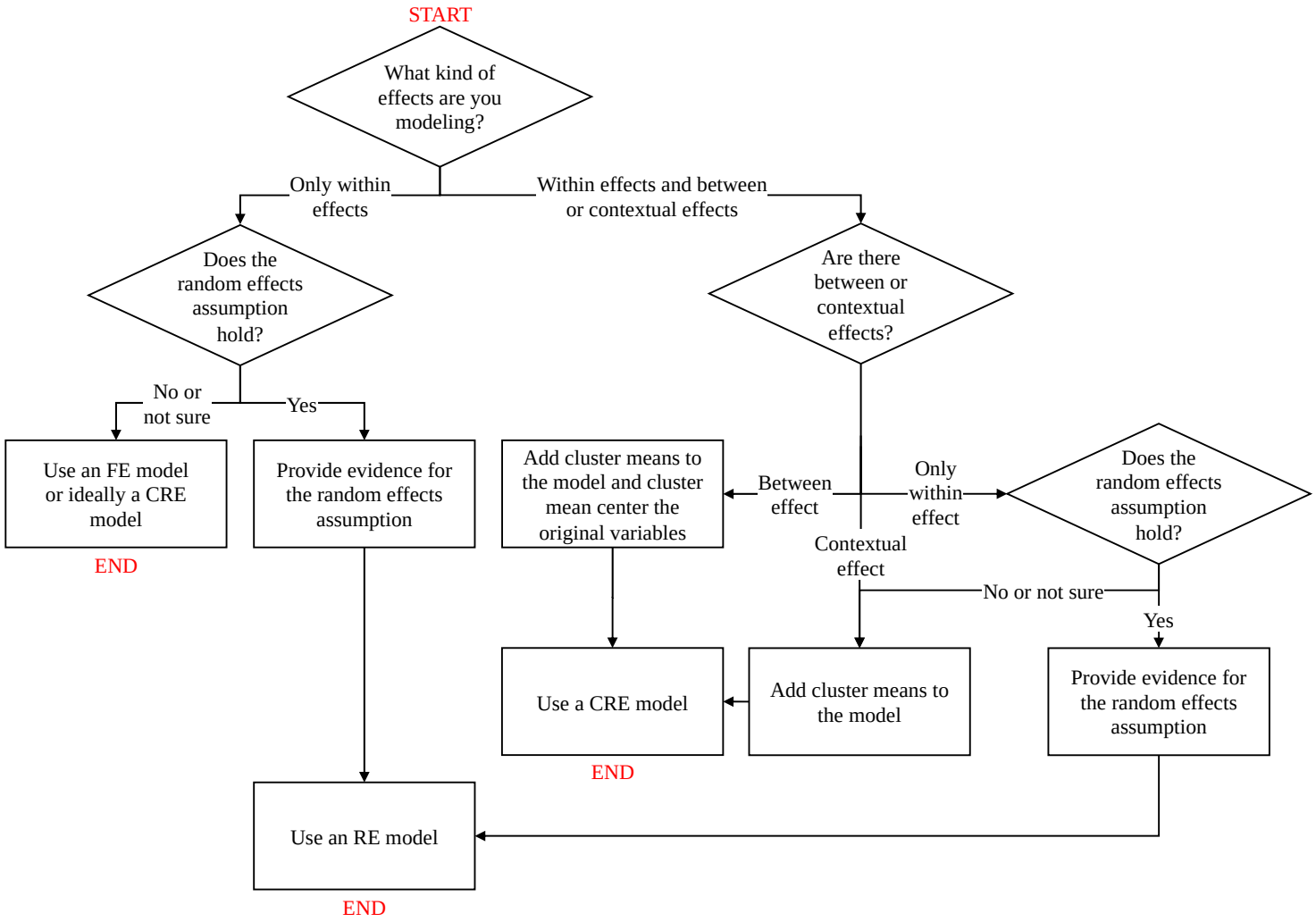


Figure 6: Proportion of articles satisfying the random effects assumption by journal



Note: FE and CRE = applied a fixed effects or correlated random effects approach that does not make the random effects assumption; RE justified = random effects approach used with appropriate justification; RE unjustified = random effects approach used without appropriate justification; Unclear = we could not ascertain what model was estimated. The above distributions did not differ significantly across journals, Fisher's exact test $p = .23$ (Pearson $\chi^2(24) = 30.56$, $p = .17$), suggesting no quality difference.

Figure 7: Decision chart to identify which estimator to use for multilevel data



Note: between or contextual effects can be also be considered each Level 1 variable separately (i.e., it is possible to include cluster means for only some of the Level 1 variables). Different ways of doing FE, RE, and CRE are explained in Table 1 and tests are explained in Table 3.

Appendix A: Video material

We include educational video material to supplement the article. The main video provides an overview of the article and its most important concepts and is available here:

https://youtu.be/mcwjto0U01I

There are also a series of shorter videos focusing on specific concepts related to multilevel models and the random-effects assumption. The full suite of videos can be accessed at

http://tiny.cc/randomeffect

Appendix B: Understanding endogeneity through omitted variable bias

To see how this bias originates, we will treat u_j as capturing all the leader-level (or in a more macro case, all firm-level) effects that may correlate both with x_{ij} and y_{ij}, but which will be omitted from the model; thus, we will look at the problem from a basic omitted variable bias point of view (cf. Antonakis, et al., 2010). The data generating model is thus:

$$y_{ij} = \gamma_0 + \gamma_1 x_{ij} + \gamma_2 u_j + e_{ij} \qquad \textit{Eq. A1}$$

As is evident above, u_j has an effect on y_{ij} as indicated by the coefficient γ_2. Now suppose we do not explicitly measure u_j because we had no idea about its effect on y_{ij} and that we estimated instead:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + v_{ij} \qquad \textit{Eq. A2}$$

The u_j term is now absorbed in v_{ij}. Will the estimate of $\beta_1 = \gamma_1$, the latter being the correct estimate? It could, but only under some restrictive conditions, as we show below. If u_j correlates with x_{ij}, irrespective of the direction of the relation, which is not relevant for the demonstration, we can model the following (and omit the intercept for expositional clarity and without a loss of generality):

$$u_j = \omega_1 x_{ij} + w_{ij} \qquad \textit{Eq. A3}$$

The endogeneity problem will become evident when substituting Eq. A3 into Eq. A1:

$$y_{ij} = \gamma_0 + \gamma_1 x_{ij} + \gamma_2(\omega_1 x_{ij} + w_{ij}) + e_{ij} \qquad \textit{Eq. A4}$$

After multiplying out we obtain the following (note what is isolated on the right hand side in parentheses):

$$y_{ij} = \gamma_0 + \gamma_1 x_{ij} + (\gamma_2 \omega_1 x_{ij} + \gamma_2 w_{ij} + e_{ij}) \qquad \textit{Eq. A5}$$

It is now obvious that the error from Eq. A2, $v_{ij} = (\gamma_2 \omega_1 x_{ij} + \gamma_2 w_{ij} + e_{ij})$, which is why we noted above that u_j is absorbed in v_{ij}. If we rearrange Eq. A5 as a function of *x* we get:

$$y_{ij} = \gamma_0 + (\gamma_1 + \gamma_2 \omega_1) x_{ij} + \gamma_2 w_{ij} + e_{ij} \qquad \textit{Eq. A6}$$

From the above, the effect of x_{ij} on y_{ij} is $\gamma_1 + \gamma_2 \omega_1$, which is the decomposed estimate of β_1 from Eq. A2. Thus, the coefficient β_1 is inconsistent because:

$$\beta_1 = \frac{Cov(y_{ij}, x_{ij})}{Var(x_{ij})} = \gamma_1 + \gamma_2 \omega_1 \neq \gamma_1 \qquad \textit{Eq. A7}$$

Thus, one of γ_2 (i.e., from Eq. A1) or ω_1 (i.e., from Eq. A3) must equal zero for β_1 to be consistently estimated and to equal γ_1.

Appendix C: Stata and R code for empirical demonstration

The code that follows show the data generation and models estimated that are presented in Table 4, using Stata and R code.

Stata code:

```
*Data generation
clear
version 15
set seed 123

*500 leaders (note: "a" is the unobserved effect)
set obs 500
gen lead_n = _n
gen a = rnormal()
gen z = round(25 + 2*rnormal() + .2*a)

*10 raters for each leader
expand 10
bys lead_n: gen rater = _n
gen x = a + rnormal()
gen y = .5*x - .5*z + 5*a + 2*rnormal()

*Data estimation

*1 OLS with Level 1 and Level 2 predictor
reg y x z, vce(cluster lead_n)

*2 OLS with cluster dummies and Level 1 predictor
reg y x i.lead_n, vce(cluster lead_n)

*3 GLS FE with Level 1 predictor
xtset lead_n rater
xtreg y x, fe vce(cluster lead_n)

*4 OLS with Level 1 predictor and its cluster mean
bys lead_n: egen x_cl = mean(x)
reg y x x_cl, cluster(lead_n)

*5 OLS with Level 1 predictor, its cluster mean, and Level 2 predictor
reg y x x_cl z, cluster(lead_n)

*6 GLS RE with Level 1 predictor, its cluster mean, and Level 2 predictor
xtreg y x x_cl z, cluster(lead_n) robust
test x_cl

*7 MLE RE with Level 1 predictor, its cluster mean, and Level 2 predictor
mixed y x x_cl z || lead:
est store one

*compare above estimate with below using likelihood ratio test
mixed y x z || lead_n:
est store two
```

```
lrtest one two

*8 ML RE with cluster-mean centered Level 1 predictor and Level 2 predictor
gen x_grp_cen = x - x_cl
mixed y x_grp_cen z || lead_n:

*9 GLS RE with cluster-mean centered Level 1 predictor and Level 2 predictor
xtreg y x z, cluster(lead_n) robust

*Wald test of the random effects assumption (equivalent to test x_cl of
*model 6)
xtoverid

*10 GLS RE with grand-mean centered Level 1 predictor and Level 2 predictor
egen x_grnd_mean = mean(x)
gen x_grnd_cen  = x - x_grnd_mean

xtreg y x_grnd_cen z, cluster(lead_n)

*11 ML RE with grand-mean centered Level 1 predictor and Level 2 predictor
mixed y x_grp_cen x_cl z || lead_n:
```

R code:

```
library(plm)
library(lme4)
library(clubSandwich)
library(lmtest)

# Data generation

set.seed(12)

# Number of leaders and raters
N <- 500
M <- 10

lead_n <- as.factor(rep(1:N,M))
a <- rep(rnorm(N),M)
z <- rep(round(25 + 2*rnorm(N) + .2*a))
rater <- rep(1:M, each = N)
x <- a + rnorm(N*M)

y <- .5*x + 5*a - .5*z + 2*rnorm(N*M)

# Data estimation

# 1 OLS with Level 1 and Level 2 predictor
m1 <- lm(y ~ x + z)
summary(m1)
coeftest(m1, vcov=vcovCR, cluster = lead_n, type ="CR2")

# 2 OLS with cluster dummies and Level 1 predictor
m2 <- lm(y ~ x + lead_n)
summary(m2)
coeftest(m2, vcov=vcovCR, cluster = lead_n, type ="CR2")
```

```
# 3 GLS FE estimator with Level 1 predictor
pdata <- pdata.frame(data.frame(y,x,z,lead_n), index = "lead_n")
m3 <- plm(y ~ x, data = pdata, model ="within")
summary(m3)
coeftest(m3, vcov=vcovCR, cluster = lead_n, type ="CR2")

# 4 OLS with Level 1 predictor and its cluster mean
x_cl <- rep(aggregate(x, list(lead_n), mean)[,2],M)
m4 <- lm(y ~ x + x_cl)
summary(m4)
coeftest(m4, vcov=vcovCR, cluster = lead_n, type ="CR2")

# 5 OLS with Level 1 predictor, its cluster mean, and Level 2 predictor
m5 <- lm(y ~ x + x_cl + z)
summary(m5)
coeftest(m5, vcov=vcovCR, cluster = lead_n, type ="CR2")

# 6 GLS RE with Level 1 predictor, its cluster mean, and Level 2 predictor
pdata <- pdata.frame(data.frame(y,x,z,x_cl,lead_n), index = "lead_n")

m6 <- plm(y ~ x + x_cl + z, data = pdata, model = "random")
summary(m6)
coeftest(m6, vcov=vcovCR, cluster = lead_n, type ="CR2")

Wald_test(m6,"x_cl", vcov="CR2", cluster = lead_n)

# MLE RE with Level 1 predictor, its cluster mean, and Level 2 predictor
m7 <- lmer(y ~ x + x_cl + z + (1|lead_n) )
summary(m7)

# compare above estimate with below using likelihood ratio test
anova(m7,lmer(y ~ x + z + (1|lead_n) ))

# 8 ML RE with cluster-mean centered Level 1 predictor and Level 2 predictor
x_grp_cen <- x - x_cl
pdata <- pdata.frame(data.frame(y,x,z,x_cl,x_grp_cen,lead_n),
                     index = "lead_n")

m8 <- lmer(y ~ x_grp_cen + x_cl + z + (1|lead_n))
summary(m8)

# GLS RE with cluster-mean centered Level 1 predictor and Level 2 predictor
m9 <- plm(y ~ x + z, data = pdata, model ="random")
summary(m9)
coeftest(m9, vcov=vcovCR, cluster = lead_n, type ="CR2")

# Wald test of the random effects assumption (equivalent to test x_cl of
# model 6)
Wald_test(update(m9, ~ . + x_cl), "x_cl", vcov="CR2", cluster = lead_n)

# 10 GLS RE with grand-mean centered Level 1 predictor and Level 2 predictor
x_grnd_cen<- x - mean(x)
pdata <- pdata.frame(data.frame(y,x,z,x_cl,x_grp_cen,x_grnd_cen,lead_n),
                     index = "lead_n")

m10 <- plm(y ~ x_grp_cen + z, data = pdata, model ="random")
```

```
summary(m10)
coeftest(m10, vcov=vcovCR, cluster = lead_n, type ="CR2")

# 11 ML RE with grand-mean centered Level 1 predictor and Level 2 predictor
m11 <- lmer(y ~ x_grnd_cen + z + (1|lead_n))
summary(m11)


library(texreg)
screenreg(list(m1,m2,m3,m4,m5,m6,m7,m8,m9,m10,m11), omit.coef ="lead_n")
```

Note: in the above, the unobserved cluster effect and the error are weighted 5 and 2 respectively (leading to a very high ICC1); for the case of the model with a lower ICC1 the unobserved cluster effect and the error are weighted 1 and 3 respectively.

Appendix D: Comparison of different estimators and centering with data having a low ICC1

Variable	(1) OLS	(2) OLS	(3) GLS FE	(4) OLS	(5) OLS	(6) GLS RE	(7) ML RE	(8) ML RE	(9) GLS RE	(10) GLS RE	(11) ML RE
x	**1.04*****	.52***	.52***	.52***	.52***	.52***	.52***		**1.04*****		
	(27.30)	(11.83)	(12.47)	(12.47)	(12.47)	(12.47)	(11.60)		(27.28)		
z	-.45***				-.46***	-.46***	-.46***	-.46***	-.45***	-.41***	-.45***
	(18.73)				(22.19)	(22.19)	(21.46)	(21.46)	(18.72)	(10.06)	(18.72)
$\bar{x}_j$				.91***	.98***	.98***	.98***	1.50***			
				(12.74)	(16.14)	(16.14)	(15.98)	(35.91)			
$x - \bar{x}_j$								.52***		.52***	
								(11.60)		(12.47)	
$x - \bar{x}$											**1.04*****
											(27.28)
Dummies		Included									
Constant	-1.34**	-14.12***	-12.58***	-12.57***	-.92*	-.92*	-.92*	-.92*	-1.34**	-2.30**	-1.35**
	(2.22)	(807.01)	(23,975.49)	(213.44)	(1.75)	(1.75)	(1.69)	(1.69)	(2.22)	(2.24)	(2.24)

Note: Cluster robust *t*-statistics in parentheses (except for ML estimator); n = 5,000 observations (clustered under N = 500 leaders). OLS = Ordinary Least Squares estimator; GLS FE = GLS fixed effects estimator; GLS RE = GLS random effects estimator; ML RE = maximum likelihood estimator with RE model; the true coefficient for x is .50 and that of age is -.50 in the population. ***$p < .01$, **$p < .05$, *$p < .10$. Note, estimates that are substantially erroneous are emphasized (bolded-underlined).

Appendix E: R code for simulations

parameters.R

```
# Simulation design matrix (full factorial)
designMatrix <- expand.grid(betax1 = -2:2,
                            betaage = -2:2,
                            l2nobs = c(20, 30, 50, 100, 500, 1000),
                            l1nobs = c(2, 5, 10, 20, 30),
                            l1endog = c(.1,.3,.5),
                            l2endog = c(0,.2,.4),
                            l1multiplier = c(1,2,4))

# Generate all possible estimators and then choose which ones are used

estimatorsToRun <- expand.grid(estimator = c("OLS","ML RE","GLS RE","GLS FE"),
                               centering =c("none","grand-mean","cluster-mean"),
                               clusterMeans = c(TRUE,FALSE),
                               stringsAsFactors = FALSE)

estimatorsToRun <- estimatorsToRun[c(1,2,3,10,11,16,18,19,22,23),]
REPLICATIONS <- 1000
```

simulations.R

```
library(MASS)
library(lme4)
library(plm)
library(plyr)

# Read the experimental conditions

source("parameters.R")


# This file is designed to be run on a computer cluster where each condition is
# run as a separate job. Read the condition number from the command line if
# given. Otherwise run all conditions and all replication sets.

args <- commandArgs(trailingOnly = TRUE)

if(length(args) == 0){
  designNumbers <- 1:nrow(designMatrix)
#  designNumbers <- 20250
} else {
  designNumbers <- as.numeric(args[1])
}


################################################################################
#
# Main program
#
################################################################################
```

```
# Loop over designs

for(designNumber in designNumbers){
  print(paste("Running design", designNumber))
  print(designMatrix[designNumber,])

  attach(designMatrix[designNumber,])

  # Generate the population correlation matrix

  C <- matrix(c(1,l2endog,l2endog,l1endog,l1endog, # uj
                l2endog,1,0,0,0,  # age
                l2endog,0,1,0,0,  # iq
                l1endog,0,0,1,.3,  # x1
                l1endog,0,0,.3,1), # x2
              5,5)

  rownames(C) <- colnames(C) <- c("uj", "age", "iq", "x1", "x2")
  counter <- 0

  set.seed(designNumber)

  reps <- replicate(REPLICATIONS, {
    counter <<- counter+1
    print(counter)

    ########################## Generate a sample ##############################

    # Generate L2 variables.
    data <- mvrnorm(l2nobs,rep(0,3),C[1:3,1:3])
    uj <- data[,1]
    age <- data[,2]
    iq <- data[,3]
    rm(data)

    lead_n <- 1:l2nobs

    # Repeate the L2 variable values so the that the lenghts match the L1
    # variable lenghts
    uj <- rep(uj,l1nobs)
    age <- rep(age,l1nobs)
    iq <- rep(iq,l1nobs)
    lead_n <- rep(lead_n,l1nobs)

    # L1 variables.
    # Regress x1 on uj, age, and iq using the population matrix and generate
    # based on predicted values
    b <- solve(C[1:3,1:3],C[1:3,4])
    r2 <- as.vector(b %*% C[1:3,1:3] %*% b)

    x1 <- (uj * b[1] +
      age * b[2] +
      iq * b[3] +
```

```
    sqrt(1-r2) * rnorm(l1nobs * l2nobs)) *
    2 * l1multiplier

  # Regress x2 on uj, age, iq, and x2 using the population matrix and generate
  # based on predicted values

  b <- solve(C[1:4,1:4],C[1:4,5])
  r2 <- as.vector(b %*% C[1:4,1:4] %*% b)

  x2 <- (uj * b[1] +
    age * b[2] +
    iq * b[3] +
    x1 * b[4] +
    sqrt(1-r2) * rnorm(l1nobs * l2nobs)) *
    2 * l1multiplier

  # Rescale all variables to have the desired means and SDs
  age <- age*2 + 30
  iq <- iq*3 + 115

  # Generate the error term
  e <- 4*rnorm(l1nobs * l2nobs) * l1multiplier

  # Generate the dependent variable
  y <- betax1*x1 + 1*x2 + betaage*age + 1*iq + 10*uj + e

  ############## Run all estimators and collect the results ################

  # Generate cluster means and different centered versions
  x1cm <- rep(aggregate(x1, list(lead_n), mean)[,2],l1nobs)
  x2cm <- rep(aggregate(x2, list(lead_n), mean)[,2],l1nobs)

  x1cc <- x1-x1cm
  x2cc <- x2-x2cm

  x1gc <- x1-mean(x1)
  x2gc <- x2-mean(x2)

  data <- data.frame(uj, age, iq, x1, x2, e, y, x1cm, x2cm, x1cc, x2cc, x1gc, x2gc,
lead_n)
  pdata <- pdata.frame(data, index = "lead_n")

  results <- matrix(NA,0,0)

  for(i in 1:nrow(estimatorsToRun)){
    f <- switch(estimatorsToRun[i,"centering"],
                none = y ~ x1 + x2 + age + iq,
                "grand mean" =  y ~ x1gc + x2gc + age + iq,
                "cluster mean" =  y ~ x1cc + x2cc + age + iq
    )

    if(estimatorsToRun[i,"clusterMeans"]=="TRUE") f <-
        update.formula(f,y ~ . + x1cm + x2cm)

    est <- switch (estimatorsToRun[i,"estimator"],
```

```
                    OLS = lm(f),
                    "ML RE" = lmer(update.formula(f,y~. + (1|lead_n)),
                                   REML = FALSE),
                    "GLS FE" = plm(f,data=pdata,model = "within"),
                    "GLS RE" = plm(f,data=pdata,model = "random")
      )

      # Extract the coefficients and variance estimates

      co <- switch(estimatorsToRun[i,"estimator"],
                   "ML RE"= fixef(est),
                   coef(est))

      va <- diag(as.matrix(vcov(est)))

      co <- rename(co,c(x1gc="x1", x1cc="x1", x2gc="x2", x2cc="x2"),
                   warn_missing = FALSE)
      va <- rename(va,c(x1gc="x1", x1cc="x1", x2gc="x2", x2cc="x2"),
                   warn_missing = FALSE)
      names(va) <- paste("var",names(va))

      a <- t(c(replication=counter,estimator=i,co,va))
      results <- rbind.fill.matrix(results,a)
    }
    results
  }, simplify = FALSE)

  results <- cbind(do.call(rbind,reps))
  save(results,file=paste("Design_",designNumber,".Rdata",sep=""))

  # Detach the design
  detach()
}
```

Appendix F: Coding Manual

1. Full article citation
2. Year of publication
3. Journal
4. Is the article focused on micro or macro questions or variables?
5. Type of panel data: Longitudinal, Hierarchical, or Both?
6. Sample size at lowest level (Level 1)?
7. Sample size at Level 2?
8. Sample size at Level 3 (if applicable)?
9. Data modeled wide or long (long format assumed if not reported)?
10. Dummies included for unobserved heterogeneity at lowest level?
11. All relevant cluster means included?
12. Estimator used (e.g., GLS random effects, GLS fixed effects, ML random effects, ML, OLS, GEE, GMM, others)?
13. Program used (e.g., HLM, Stata, Mplus, SPSS, R)?
14. Command used if reported by authors?
15. Are modeled Level 1 predictors endogenous, exogenous (i.e., manipulated, fixed, or vary randomly in nature), or instrumented?
16. Are modeled Level 2 predictors endogenous, exogenous (i.e., manipulated, fixed, or vary randomly in nature), or instrumented?
17. Are Level 1 data centered (i.e., grand-mean, cluster-mean, standardized)?
18. Are Level 2 data centered (i.e., grand-mean, cluster-mean, standardized)?
19. Are authors interested in Level 1 effects, Level 2 effects, both Level 1 & Level 2, or cross-level interactions?
20. Is the random effects assumption (if applicable) tested for?
21. If test of the RE assumption, which test is performed (e.g., Hausman, LR, F-test)?
22. If test of the RE assumption, does the RE assumption empirically hold?
23. Do authors make the RE assumption, and if yes, does it empirically hold?
24. How are standard errors computed (e.g., default, heteroscedasticity robust, cluster robust)
25. Do authors model time effects in longitudinal studies (if applicable)?

26. Do authors make causal claims, recognize correlation, acknowledge some causality issues in the limitations, or are unclear regarding causality

Appendix G: Coding variables across journals

	AMJ (N=44)	*SMJ (N=38)*	*JAP (N=27)*	*JOM (N=22)*	*JOB (N=19)*	*OS (N=16)*	*LQ (N=14)*	*PP (N=13)*	*JOpM (N=11)*	***All journals (N=204)***
Focus										
Micro	45.45%	0%	96.30%	59.09%	94.74%	0%	85.71%	100%	0%	**50%**
Macro	54.55%	100%	3.70%	40.91%	5.26%	100%	14.29%	0%	100%	**50%**
Panel Type										
Longitudinal	59.09%	57.89%	37.04%	40.91%	47.37%	56.25%	21.43%	30.77%	81.82%	**46.67%**
Hierarchical	25.00%	23.68%	59.26%	54.55%	42.11%	18.75%	78.57%	69.23%	9.09%	**45.33%**
Both longitudinal & hierarchical	15.91%	18.42%	3.70%	4.55%	10.53%	25.00%	0%	0%	9.09%	**8.00%**
Median Level 1 Sample size										
Longitudinal	2,012	1,496	885	1,610	655	8,847	685	592	5,033	**1,184**
Hierarchical	525	2,115	661	300	361	3,032	291	213	187	**328**
Both longitudinal & hierarchical	17,658	51,730	2920	483	11,885	7,006	N/A	N/A	878	**7,488**
Median Level 2 Sample size										
Longitudinal	240	141	93	299	86	482	137	75	304	**150**
Hierarchical	68	122	65	45	59	557	63	54	12	**58**
Both longitudinal & hierarchical	2,332	6,692	584	N/R	2,916	517	N/A	N/A	307	**2,155**

Note: N/A: Not applicable; N/R: Not reported or unclear. AMJ=Academy of Management Journal; SMJ=Strategic Management Journal; JAP=Journal of Applied Psychology; JOM=Journal of Management; JOB=Journal of Organizational Behavior; OS=Organization Science; LQ = Leadership Quarterly; PP=Personnel Psychology; JOpM=Journal of Operations Management. Note, overall, 83.17% of the articles analyzed data at two levels, and the rest had three levels of data.

Appendix H: Coding categories for different variables under study

Estimator used	*Program used*	*Interest*	*Standard error*	*Treatment of time*	*Test of random effects*
ML random effects (106)	HLM (43)	Level 1 effects (118)	Nothing reported about standard errors, assuming default (113)	Time dummies (85)	No tests performed despite the use of RE (90)
OLS, 2SLS, or 3SLS (32)	Mplus (32)	Level 2 effects (4)	Cluster robust (57)	Ignored (37)	Not applicable because no RE model used (80)
ML non random effects (29)	Stata (30)	Interest in Level 1 and Level 2 effects (22)	Heteroscedastic robust SE (28)	Not applicable (82)	Hausman test (24)
GLS random effects (12)	SAS (5)	Interest in Level 1/Level 2 effects and in cross level interactions (60)	Others (6)		Comparison with other estimation procedures, but no test (10)
GLS fixed effects (10)	R (4)				
GMM (7)	SPSS (2)				
GEE (4)	Others, unclear and not-reported (88)				
Others, unclear and not-reported (4)					

Note: Number in parentheses represents the number of articles which have taken the following approach.