Journal of Management
Vol. 39 No. 6, September 2013 1490-1528
DOI: 10.1177/0149206313478188

Best-Practice Recommendations for Estimating Cross-Level Interaction Effects Using Multilevel Modeling

Herman Aguinis
Ryan K. Gottfredson
Indiana University
Steven Andrew Culpepper
University of Illinois at Urbana-Champaign

Multilevel modeling allows researchers to understand whether relationships between lower-level variables (e.g., individual job satisfaction and individual performance, firm capabilities and performance) change as a function of higher-order moderator variables (e.g., leadership climate, market-based conditions). We describe how to estimate such cross-level interaction effects and distill the technical literature for a general readership of management researchers, including a description of the multilevel model building process and an illustration of analyses and results with a data set grounded in substantive theory. In addition, we provide 10 specific best-practice recommendations regarding persistent and important challenges that researchers face before and after data collection to improve the accuracy of substantive conclusions involving cross-level interaction effects. Our recommendations provide guidance on how to define the cross-level interaction effect, compute statistical power and make research design decisions, test hypotheses with various types of moderator variables (e.g., continuous, categorical), rescale (i.e., center) predictors, graph the cross-level interaction effect, interpret interactions given the symmetrical nature of such effects, test multiple cross-level interaction hypotheses, test cross-level interactions involving more than two levels of nesting, compute effect-size estimates and interpret the practical importance of a cross-level interaction effect, and report results regarding the multilevel model building process.

Keywords: *multilevel modeling; moderation; cross-level; interaction*

Acknowledgments: We thank Fred Oswald and three Journal of Management *anonymous reviewers for highly constructive feedback that allowed us to improve our article substantially. Also, we thank the more than 2,000 members of the listservs RMNET (Research Methods Division of the Academy of Management) and MULTILEVEL for providing us with questions and challenges that we addressed in our article.*

Corresponding author: Herman Aguinis, Department of Management and Entrepreneurship, Kelley School of Business, Indiana University, 1309 E. 10th Street, Bloomington, IN 47405-1701, USA.

E-mail: haguinis@indiana.edu

Integrating micro and macro levels of analysis is one of the biggest challenges in the field of management (Aguinis, Boyd, Pierce, & Short, 2011). Specifically, there is an interest in integrating theories that explain and predict phenomena at the individual, team, and organizational levels of analysis (Bliese, 2000; Kozlowski & Klein, 2000; Liden & Antonakis, 2009; Mathieu & Chen, 2011; Molloy, Ployhart, & Wright, 2011). When conducting research that includes variables measured at different levels of analysis, researchers explicitly recognize that lower-level entities such as individuals are nested within higher-level collectives such as teams. Note that lower-level entities do not have to be individuals. For example, lower-level entities can be organizations and higher-level collectives can be industries, countries, or economic blocks (e.g., MERCOSUR, European Union). Regardless of the specific definition of entities and the collectives within which they reside, the multilevel nature of the resulting data requires that dependence among observations be considered both conceptually and analytically (Snijders & Bosker, 2012). Of particular interest in terms of integrating micro and macro domains is whether the nature of a lower-level relationship depends on a higher-level factor—what we label a *cross-level interaction effect*. Conceptually, there is a need to consider theoretical reasons for expecting a cross-level interaction effect, and, analytically, the resulting data should be examined using appropriate tools.

Dependence is not solely a function of whether observations are formally clustered into larger units. As noted by Kenny and Judd (1996: 138), “[O]bservations may be dependent, for instance, because they share some common feature, come from some common source, are affected by social interaction, or are arranged spatially or sequentially in time.” Thus, dependence of observations also occurs when shared experiences and context affect lower-level units such as firms in the same industry facing similar market-based challenges, different branches of a bank being influenced by the same strategic priorities established for a particular geographic region, or employees within a team being similarly affected by the ineffective communication style of their supervisor. In other words, a higher-level variable may covary with relevant lower-level outcome variables, and entities within collectives may be more similar regarding certain variables compared to entities across collectives (Bliese & Hanges, 2004). Consequently, dependence may occur “even if the variable of interest makes no reference to the group” (Bliese, 2000: 358). Covariation between higher-level variables and lower-level outcomes leads to gross errors of prediction if a researcher uses statistical approaches such as ordinary least squares (OLS) regression, which are not designed to model data structures that include dependence due to clustering of entities (Bliese & Hanges, 2004; Hox, 2010; Snijders & Bosker, 2012).

Although moderated multiple regression (MMR) is arguably the most popular data-analytic approach for estimating interaction effects in management and related fields (Aguinis, Beaty, Boik, & Pierce, 2005), it is highly impractical in the presence of nested data structures (Davison, Kwak, Seo, & Choi, 2002). Moreover, although MMR could be used to understand whether situations conceptualized as categorical groupings or conditions interact with lower-level predictors, MMR forces the situation to be conceptualized as categorical differences (or “treatments”). Alternatively, a multilevel analytical approach allows for an investigation of influences, both direct and interactive, of continuous higher-level variables on lower-level outcomes (Mathieu, Aguinis, Culpepper, & Chen, 2012). So, multilevel modeling offers a practical as well as substantive advantage regarding the estimation of cross-level interaction effects compared to MMR.

Much of the literature on multilevel modeling is quite specialized, including analytic work that is mathematically sophisticated as well as Monte Carlo simulations involving lengthy and complex procedures and results. Due to the nature of this research, much of this work is not easily accessible to researchers with the usual methodological and statistical background obtained from doctoral-level training in management and related fields. Accordingly, our article distills the technical literature for a general readership and includes 10 specific best-practice recommendations that researchers will be able to implement in their own quest for interaction effects involving variables at different levels of analysis. Our article makes a dual contribution. First, it offers a “one-stop-shopping experience” regarding multilevel modeling analysis in general, and, second, it also offers specific recommendations regarding the test and interpretation of cross-level interactions in particular. Regarding our article’s first contribution, we rely on several excellent books available (e.g., Hox, 2010; Kreft & De Leeuw, 1998; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012), but we provide quicker access to useful recommendations and tools, specific examples and actionable pointers, and a less technical treatment that, taken together, make the material accessible to a wider audience of researchers in management and related fields.

Next, we provide a conceptual and technical description of the steps involved in estimating cross-level interaction effects. The technical aspects of our presentation are necessary because they provide the foundation for the best-practice recommendations we offer in the following section. Although some of the material in the next section involves formulae with which some readers may not be very familiar, we provide an explanation of each and also accompany them with illustrations based on a realistic research situation. Moreover, we also offer graphs to enhance the pedagogical value of the technical material.

Estimating Cross-Level Interaction Effects Using Multilevel Modeling

We created a data file including $N = 630$ individuals nested in $J = 105$ teams patterned after a study by Chen, Kirkman, Kanfer, Allen, and Rosen (2007) to provide a realistic scenario grounded in substantive theory. Specifically, Chen et al. investigated whether the quality of leader–member exchange (LMX) (X) predicts individual empowerment (Y) given data collected across teams that differ regarding leadership climate (W), and all three variables were measured using 7-point Likert-type scales. Overall, Chen et al.’s theoretical model predicted that employees who report higher LMX (i.e., a better relationship with their leader) will feel more empowered (i.e., they have the autonomy and capability to perform meaningful work that can affect their organization). In addition, Chen et al.’s model included the hypothesis that the team-level variable leadership climate (i.e., ambient leadership behaviors directed at the team as a whole) would also affect individual-level empowerment positively. Moreover, Chen et al. hypothesized that the relationship between LMX and empowerment would be moderated by leadership climate such that the relationship would be stronger for teams with a better leadership climate. The data file and the annotated R code used to conduct all the analyses described in our article are available at http://mypage.iu.edu/~haguinis. The annotated R code is also included in Appendix A. The availability of the data file and R code will allow readers to replicate the illustrative analyses and results we describe throughout our article.

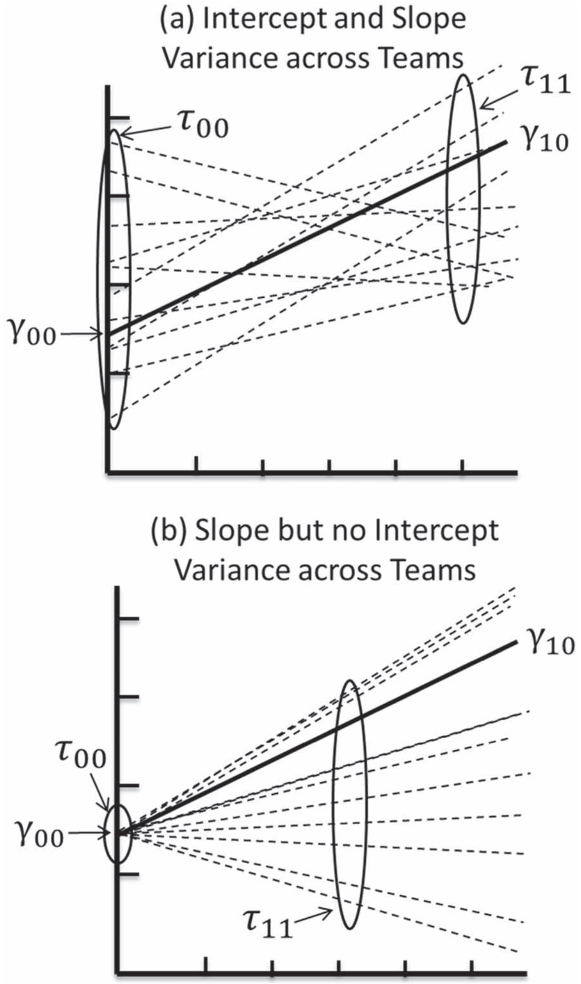
In the context of multilevel modeling, it is possible to test hypotheses regarding three types of relationships or effects (note that for ease of presentation we use the term *effect* in the remainder of our article although in some studies causal relationships may not be clearly established due to the use of nonexperimental designs):

1. *Lower-level direct effects*. Does a lower-level predictor *X* (i.e., Level 1 or L1 predictor) have an effect on a lower-level outcome variable *Y* (i.e., L1 outcome)? Specifically regarding our illustration, there is an interest in testing whether LMX, as perceived by subordinates, predicts individual empowerment. Note that LMX scores are collected for each individual worker (i.e., there is no aggregation of such scores for the purpose of testing the presence of a lower-level direct effect).
2. *Cross-level direct effects*. Does a higher-level predictor *W* (i.e., Level 2 or L2 predictor) have an effect on an L1 outcome variable *Y*? Specifically, we would like to assess whether L2 variable leadership climate predicts L1 outcome individual empowerment.
3. *Cross-level interaction effects*. Does the nature or strength of the relationship between two lower-level variables (e.g., L1 predictor *X* and L1 outcome *Y*) change as a function of a higher-level variable *W*? Referring back to our substantive illustration, we are interested in testing the hypothesis that the relationship between LMX and individual empowerment may vary as a function of (i.e., is moderated by) the degree of leadership climate such that the relationship will be stronger for teams with more positive leadership climate and weaker for teams with less positive leadership climate.

Although our article's specific goal is to discuss issues about cross-level interaction effects, as noted above, researchers using multilevel modeling are usually interested in assessing other effects as well. Overall, there is an interest in understanding factors that may explain three key sources of variance that parallel the three types of effects we just described: (1) What are the L1 factors that explain within-group variance (i.e., lower-level direct effects)? (2) What are the L2 factors that explain across-group variance in intercepts (i.e., cross-level direct effects)? and (3) What are the group-level factors that explain variance in across-group slopes (i.e., cross-level interaction effects)? These same three questions are the focus of multilevel analyses regardless of the nature of the constructs and the particular measurement approach adopted to measure them (e.g., multiple-indicator measures, multidimensional constructs; Preacher, Zyphur, & Zhang, 2010).

To enhance the clarity of our presentation, we offer a visual representation of these three sources of variance—we will provide a more detailed analytic treatment after the graphical descriptions. The dashed lines in Figure 1's top and bottom panels show that we can estimate an OLS regression equation for the relationship between LMX and empowerment within each team. Thus, each team has its own regression line defined by its own intercept and slope. Figure 1's panels also show a solid line, which is a pooled regression line between LMX scores and empowerment across all teams. This pooled regression line is defined by its own intercept (i.e., γ_{00}; "gamma sub zero zero") and slope (i.e., γ_{10}; "gamma sub one zero"). Figure 1's Panel (a) also shows that regression lines differ across teams in terms of both intercepts and slopes. As shown in Figure 1's Panel (a), the variance of intercepts across teams is denoted by τ_{00} ("tau sub zero zero") and the variance of slopes across teams is denoted by τ_{11} ("tau sub one one"). In contrast, illustrating a different yet possible research

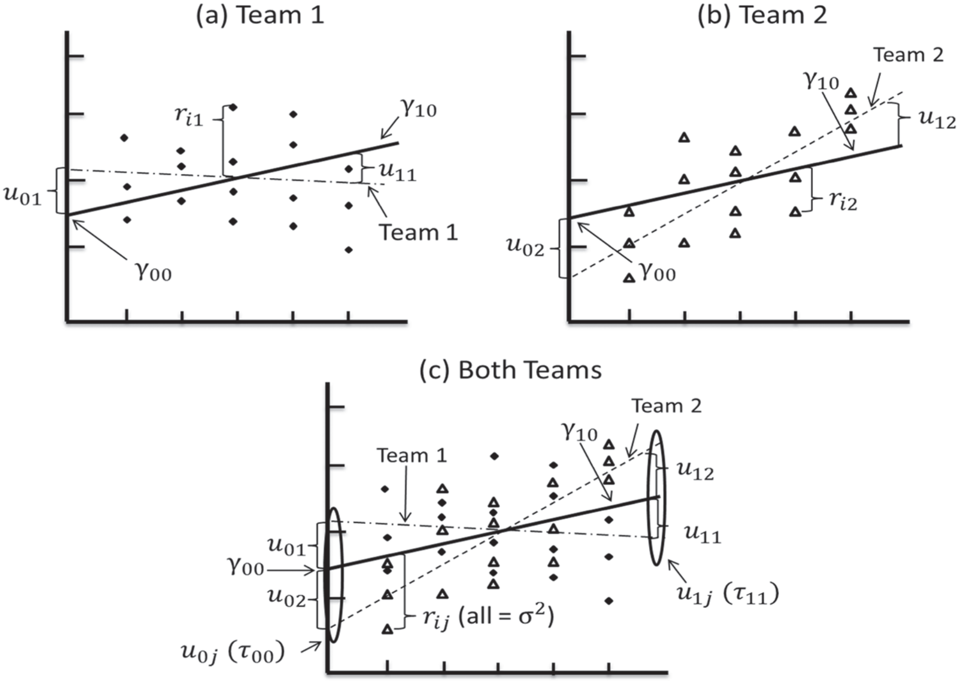
Figure 1
Illustration of Variance of Intercepts (τ_{00}) and Slopes (τ_{11}) Across Teams



scenario, Figure 1’s Panel (b) shows that teams differ regarding slopes (i.e., $\tau_{11} > 0$), but not regarding intercepts (i.e., $\tau_{00} = 0$).

Figure 2 includes a graphic depiction of individual data points within two teams only: Team 1 in Panel (a) and Team 2 in Panel (b). Figure 2’s Panel (c) shows data for all individuals from both of these teams combined. Similar to Figure 1, Figure 2’s Panel (a) shows the OLS regression line for Team 1 (dashed line) as well as the pooled regression line for all teams (solid line). Also, Panel (a) shows the L1 residual or error scores r_{i1} (i.e., differences between observed and predicted score for empowerment based on LMX scores within Team 1).

Figure 2
Illustration of Within-Group Variance (σ^2), and Level 2 Intercept Variance (τ_{00}) and Slope Variance (τ_{11}) Across Teams



Note that the variance of these residual scores within teams is symbolized by σ^2 in Figure 2's Panel (c). Figure 2's Panel (a) also shows the difference between the Team 1 intercept and the pooled (across all teams) intercept γ_{00} (i.e., L2 residual), which is symbolized by u_{01}. In addition, Panel (a) shows the difference between the Team 1 slope and the pooled (across all teams) slope γ_{10} (i.e., L2 residual), which is symbolized by u_{11}. That is u_{11} is nonzero when Team 1's prediction equation has a different slope than the pooled line. Similarly, Panel (b) also shows the OLS regression line for this particular team, the pooled regression line (across all teams), and the L1 and L2 residuals.

Panel (c) in Figure 2 shows individuals from Team 1 and Team 2 combined. For clarity, this panel includes only Team 1 and Team 2 from the many teams in Figure 1. This panel shows the three sources of variance that we are interested in understanding using multilevel modeling: variance of the L1 residuals r_{ij} (i.e., σ^2, within-group variance), variance of the L2 residuals u_{0j} (i.e.,τ_{00}, intercept variance across teams), and variance of the L2 residuals u_{1j} (i.e.,τ_{11}, slope variance across teams).

Analytically, Figure 2's Panel (c) can be described by the following L1 model (Raudenbush & Bryk, 2002; Singer, 1998):

$$Y_{ij} = \beta_{0j} + \beta_{1j}\left(X_{ij} - \overline{X}_j\right) + r_{ij} \tag{1}$$

Equation 1 takes on the familiar OLS regression equation form because it includes a predictor and a criterion residing at the same level of analysis (i.e., L1 in this case). Specifically, Y_{ij} is the predicted empowerment score for the ith person in team j, β_{0j} is the intercept parameter for team j, β_{1j} is the slope parameter for team j, X_{ij} is the individual LMX for the ith person in team j and is rescaled (i.e., "centered") by the team average $\overline{X}_j$. As discussed later in our article, this type of rescaling, called "group mean-centering" or "within-cluster centering," is one of two approaches available. The term r_{ij} is the L1 residual term (i.e., randomly distributed error), reflecting individual-level differences in empowerment around the predicted empowerment score for employees within each team. As mentioned earlier, our interest does not focus on the residual scores per se, but in the variance of r_{ij}, denoted by σ^2, which represents the amount of within-group variance for the criterion scores (i.e., individual empowerment). Note that σ^2 is analogous to $\text{MS}_{\text{within}}$ in analysis of variance (ANOVA), and, as discussed earlier, it is illustrated graphically in Figure 2's Panel (c).

The interpretation of the parameter β_{0j} depends on the scaling of the predictor X_{ij}. To establish a meaningful interpretation of this parameter, Equation 1 rescales the predictor by each team's mean. Consequently, the mean of $X_{ij} - \overline{X}_j$ is zero within teams and β_{0j} is interpreted as the predicted level of empowerment for a typical (i.e., mean) LMX of members of a given team. Note that instead of rescaling by the group mean, we could rescale X_{ij} by any other value for LMX, say 4.5 on a 5-point scale. So, if we rescale by 4.5, β_{0j} would be interpreted as the predicted level of empowerment for individuals in a given team with a team average LMX score of 4.5. Finally, based on Equation 1, the parameter β_{1j} is interpreted as the predicted increase in individual empowerment associated with a 1-unit increase in LMX for individuals within the j^{th} team.

The multilevel model building process usually involves a sequence including four steps. The first step involves what is labeled an *unconditional means, one-way random-effects ANOVA*, or *null model*. The second step involves what is called a *random intercept and fixed slope model*. The third step involves the *random intercept and slope model*. Finally, the fourth step involves the *cross-level interaction model*. Although our best-practice recommendations are particularly relevant to the third and fourth steps, next we provide a description of each of the steps involved in the model building process.

Step 1: Null Model

The *null model* begins with specifying the following relationship,

$$\text{Null model (Level 1)}: Y_{ij} = \beta_{0j} + r_{ij}, \tag{2}$$

which is identical to Equation 1 but excludes the L1 predictor. Due to the nested nature of the data, it is possible that both the intercept and slope in Equation 1 vary across teams. Specifically, it is likely that teams differ in average empowerment (i.e., β_{0j} differs across the J teams) and individual team members' LMX levels may relate differently to empowerment across teams (i.e., β_{1j} differs across the J teams). This situation is illustrated in Panel (a) of Figure 1. However, in this first step in the model building process, we omit predictors and only allow intercepts to vary across teams. Formally stated,

$$\text{Null model (Level 2): } \beta_{0j} = \gamma_{00} + u_{0j}, \tag{3}$$

In Equation 3, the team intercepts are shown to be a function of the grand mean (i.e., averaged across all teams) intercept γ_{00} and a residual term u_{0j} that describes how team intercepts deviate from the grand mean intercept. Substituting Equation 3 into Equation 2 leads to the following combined model:

$$\text{Null model (Combined): } Y_{ij} = \gamma_{00} + u_{0j} + r_{ij} \tag{4}$$

Referring back to our substantive illustration, the combined null model in Equation 4 shows that individual empowerment is a function of the grand mean LMX (i.e., γ_{00}) across-group differences in individual empowerment scores (i.e., L2 residuals u_{0j}), and within-group differences in individual empowerment scores (i.e., L1 residuals r_{ij}). As noted earlier, the variance of u_{0j}, denoted by τ_{00}, quantifies the degree of heterogeneity in intercepts across teams and the variance of r_{ij}, denoted by σ^2, quantifies the within-group variance. Thus, in comparison to an ANOVA framework, σ^2 is analogous to MS_{within} and τ_{00} is analogous to $MS_{between}$. In terms of our illustration, τ_{00} quantifies the variation in mean empowerment scores across teams. A key difference between ANOVA and multilevel modeling, however, is that multilevel modeling conceptualizes the teams as a random sample from a larger population of teams (i.e., a random factor), whereas ANOVA conceptualizes the teams as being qualitatively different (i.e., a fixed factor). Furthermore, as we will reiterate later in our article, the label *fixed effects* is reserved for multilevel modeling estimates that are constant across L2 units, such as γ_{00}, and the label *random effects* is used to denote the model estimates that vary across L2 units (e.g., u_{0j}).

As part of the first step in the model building process, we compute the intraclass correlation (ICC), which quantifies the proportion of the total variation in individual empowerment accounted for by team differences. An alternative interpretation is that the ICC is the expected correlation between empowerment scores for two individuals who are in the same team. ICC = $\tau_{00} / [\tau_{00} + \sigma^2]$ and it ranges from 0 to 1. A value near zero suggests that a model including L1 variables only is appropriate, and, hence, there may be no need to use multilevel modeling. Instead, a simpler OLS regression approach may be more parsimonious. On the other hand, ICC > 0, even as small as .10 (Kahn, 2011), suggests that there may be a L2 variable W (e.g., leadership climate) that explains heterogeneity of empowerment scores across teams (i.e., β_{0j}). Moreover, OLS standard errors and significance tests may be compromised in the presence of even smaller ICCs. Based on a review of articles published in *Journal of Applied Psychology* between 2000 and 2010, Mathieu et al. (2012) found that ICC values reported in multilevel studies usually range from .15 to .30. Similarly, based on the educational literature, Hedges and Hedberg (2007) concluded that ICC values typically range from .10 and .25, and, based on the school psychology literature, Peugh (2010) reported a range of ICCs from .05 to .20. These results may be indicative that higher-level

Table 1
Results of Multilevel Modeling Analysis With Illustrative Data

	Model			
Level and Variable	Null (Step 1)	Random Intercept and Fixed Slope (Step 2)	Random Intercept and Random Slope (Step 3)	Cross-Level Interaction (Step 4)
Level 1				
Intercept (γ_{00})	5.720** (0.045)	5.720** (0.038)	5.720** (0.038)	5.720** (0.038)
LMX (γ_{10})		0.279** (0.023)	0.270** (0.028)	0.269** (0.027)
Level 2				
Leadership climate (γ_{01})		0.351** (0.055)	0.356** (0.055)	0.351** (0.055)
Cross-level interaction				
LMX leadership climate (γ_{11})				0.104** (0.037)
Variance components				
Within-team (L1) variance (σ^2)	0.714	0.563	0.514	0.516
Intercept (L2) variance (τ_{00})	0.095	0.060	0.068	0.068
Slope (L2) variance (τ_{11})			0.025	0.019
Intercept-slope (L2) covariance (τ_{01})			0.004	–0.004
Additional information				
ICC	0.117			
–2 log likelihood (FIML)	1,637	1,478**	1,469*	1,462**
Number of estimated parameters	3	5	7	8
Pseudo R^2	0	.23	.23	.24

Note: FIML = full information maximum likelihood estimation; L1 = Level 1; L2 = Level 2. L1 N = 630 and L2 sample size = 105. Values in parentheses are standard errors; t-statistics were computed as the ratio of each regression coefficient divided by its standard error. The data and annotated R code used for producing the results reported in this table are available at http://mypage.iu.edu/~haguinis.
$*p < .05$. $**p < .01$.

influences are more common than typically assumed, or even considered, in management research (Liden & Antonakis, 2009).

We conducted analyses pertaining to the first step in the model building process using our illustrative data file. Results included in Table 1 indicate that ICC = .117, which means that differences across teams account for about 11.7% of the variability in individuals' empowerment levels. As shown in Table 1, the across-team variance in individual empowerment is τ_{00} = .095 and the within-team variance is .714. In short, results provide evidence for a nested data structure that requires multilevel modeling rather than a single-level data analytic approach. Table 1 also shows additional results pertaining to the combined null model. We will describe the interpretation of all of the results included in Table 1 in subsequent sections of our article.

Step 2: Random Intercept and Fixed Slope Model

As a second step in the model building process, we may be interested in understanding the factors that explain σ^2 and τ_{00}. This second step involves creating what is labeled a *random intercept and fixed slope model* (RIFSM), which begins with the following equation,

$$\text{RIFSM (Level 1):}\, Y_{ij} = \beta_{0j} + \beta_{1j}\left(X_{ij} - \overline{X}_j\right) + r_{ij}, \tag{5}$$

which is identical to Equation 1. The next step in the process of building the RIFSM involves adding the L2 equations as follows (Enders & Tofighi, 2007; Hofmann & Gavin, 1998),[1]

$$\text{RIFSM (Level 2):}\ \beta_{0j} = \gamma_{00} + \gamma_{01}\left(W_j - \overline{W}\right) + u_{0j}, \tag{6}$$

where the team intercepts are shown to be a function of the grand mean (i.e., averaged across all teams) intercept γ_{00} and a residual term u_{0j} that describes how teams deviate from the grand mean, after controlling for team leadership. Also, γ_{01} is interpreted as the amount of change in a team's average empowerment score associated with a 1-unit increase in leadership climate. In this model the slopes are not allowed to vary across teams and, hence,

$$\text{RIFSM (Level 2):}\, \beta_{1j} = \gamma_{10}\ , \tag{7}$$

which leads to the following combined model:

$$\text{RIFSM (Combined):}\ Y_{ij} = \gamma_{00} + \gamma_{10}\left(X_{ij} - \overline{X}_j\right) + \gamma_{01}\left(W_j - \overline{W}\right) + u_{0j} + r_{ij} \tag{8}$$

Note that Equation 8 is called a RIFSM because it allows intercepts (i.e., mean scores) to vary across teams by the inclusion of u_{0j}. However, as shown in Equation 7, slopes are not allowed to vary across teams. Rather, as shown in Equation 8, one fixed value for the slope of empowerment on LMX scores (i.e., γ_{10}) is used for all individuals regardless of team membership. In other words, the relationship between LMX and empowerment is assumed to be identical across all teams (similar to the assumption in ANCOVA; Culpepper & Aguinis, 2011). In sum, Equation 8 predicts individual empowerment scores based on a common intercept, γ_{00}, individual LMX scores (L1 predictor) reflected by the coefficient γ_{10}, and leadership climate (L2 predictor) reflected by the coefficient γ_{01}. In other words, γ_{01} assesses the possible presence of a cross-level direct effect (i.e., effect of leadership climate on individual empowerment) controlling for individual-level LMX scores and, therefore, explains at least part of τ_{00} identified in the first step of the model building process.

In Equation 8, γ_{00} represents mean empowerment for a team with a leadership climate score at the mean $\overline{W}$, γ_{01} is the amount of change in a team's average empowerment score associated with a 1-unit increase in leadership climate, and u_{0j} is a residual term (i.e., errors) in predicting teams' average empowerment after controlling for L2 variable leadership climate. Note that W_j is rescaled by the average team leadership climate ($\overline{W}$) to interpret γ_{00} in reference to $\overline{W}$. As was the case in Equation 1, we can rescale W_j using other values, which would lead to a different interpretation for γ_{00}. In our particular situation, γ_{01} is the predicted slope for regressing empowerment on leadership climate for teams with a mean leadership climate score of $\overline{W}$.

Once again, we used our illustrative data file and the annotated R code to produce results pertaining to this second step in the model building process. Note that the data file includes the raw (i.e., original) as well as rescaled (i.e., centered) scores. As described earlier, we used rescaled scores for our analyses. As shown in Table 1, results indicate that mean empowerment for a team with a leadership climate score at the mean $\overline{W}$ is $\gamma_{00} = 5.72$. Table 1 also shows that a 1-unit increase in leadership climate is associated with a $\gamma_{01} = .351$ increase in a team's average empowerment score. Also, Table 1 shows that the predicted slope regressing empowerment on LMX is $\gamma_{10} = .279$. In short, results provide evidence in support of a direct single-level effect (i.e., individual LMX on individual empowerment) as well as a direct cross-level effect (i.e., team-level leadership climate on individual-level empowerment).

Step 3: Random Intercept and Random Slope Model

As a third step in the model building process, we are interested in understanding whether the third key source of variance, the variance of slopes across groups (i.e., τ_{11}), is different from zero. In other words, we would like to answer the question of whether the relationship between LMX scores and empowerment varies across teams. There is no point in examining which particular moderators may explain slope variance across teams if such variance is nonexistent. To do so, we build a *random intercept and random slope model* (RIRSM) that adds a random slope component so that β_{1j} is allowed to vary across teams.

First, as usual, we begin the model building process with the L1 equation (identical to Equation 1):

$$\text{RIRSM (Level 1): } Y_{ij} = \beta_{0j} + \beta_{1j}\left(X_{ij} - \overline{X}_j\right) + r_{ij}\,. \tag{9}$$

Then, we allow both intercepts and slopes to vary across teams as follows:

$$\text{RIRSM (Level 2): } \beta_{0j} = \gamma_{00} + \gamma_{01}\left(W_j - \overline{W}\right) + u_{0j} \tag{10}$$

$$\text{RIRSM (Level 2): } \beta_{1j} = \gamma_{10} + u_{1j}. \tag{11}$$

In Equation 11, the slope of empowerment on LMX scores is a function of the grand mean (i.e., estimated across all teams) slope γ_{10} and a residual term u_{1j} that describes how team slopes differ from the pooled slope across teams. Substituting Equations 10 and 11 into Equation 9 yields the combined RIRSM as follows:

$$\text{RIRSM (Combined): } Y_{ij} = \gamma_{00} + \gamma_{01}\left(W_j - \overline{W}\right) + \gamma_{10}\left(X_{ij} - \overline{X}_j\right) + u_{0j} + u_{1j}\left(X_{ij} - \overline{X}_j\right) + r_{ij} \tag{12}$$

A comparison of the combined RIFSM (Equation 8) with the RIRSM (Equation 12) seems to suggest that the only difference is that in the latter we allow the slope of empowerment on LMX to vary across teams by the inclusion of u_{1j} and its variance τ_{11}. However, there is one additional parameter estimate that is not explicit in the model: the covariance between intercepts and slopes, which is denoted by τ_{01}. Thus, the RIRSM includes two parameters that are not part of the RIFSM: τ_{11} and τ_{01}. Referring back to our substantive illustration, a positive value of τ_{01} means that teams with steeper slopes (i.e., stronger relationship) of empowerment on LMX tend to have higher team empowerment levels.

Based on Equation 12, we can examine the standard error estimate, which is standard output in most software packages such as HLM and SAS, to answer the question of whether the variance of the residuals u_{1j} (i.e., τ_{11}) is nonzero (Bliese, 2002). Specifically, the output file in some software packages includes a confidence interval, computed based on the standard error, for the estimate of τ_{11}. If the lower bound does not include zero, then we conclude that the slope of empowerment on LMX scores varies across teams. However, in spite of its availability in many software packages, creating a confidence interval around τ_{11} can lead to incorrect conclusions. There are two reasons for this. First, standard errors for the variance components of the model, such as τ_{11}, are usually inaccurate. As concluded by Maas and Hox (2004: 437) based on an extensive simulation study, "The estimates of the variances are unbiased, but the standard errors are not always accurate." Second, a confidence interval is created by adding and subtracting the same value, such as 1.96 for a 95% interval, and, therefore, the assumption is that the parameter estimate is normally distributed, which is "doubtful for estimated variances; for example, because these are necessarily nonnegative" (Snijders & Bosker, 2012: 100). Accordingly, a better alternative for creating a confidence interval around τ_{11} is to implement a nonparametric residual bootstrap procedure as described by Carpenter, Goldstein, and Rasbash (2003). Our recommendation regarding the use of this type of confidence interval is also supported by theoretical evidence regarding its accuracy as described by Field and Welsh (2007).

A second option in terms of understanding whether τ_{11} is different from zero is to compute a −2 log likelihood ratio between Equation 12 (i.e., model with a random slope component) and Equation 8 (i.e., model without a random slope component; Bliese, 2002). A log-likelihood value quantifies the probability that the model being estimated produced the sample data (Peugh, 2010). Multiplying the log likelihood value by −2 yields a value labeled "deviance," which can be used to compare the relative fit of two competing models. Note that, when full information maximum likelihood (FIML) is used, the deviance value shows how well the variance-covariance estimates (i.e., τ_{00}, τ_{01}, and τ_{11}) *and* the regression coefficients fit the sample data. However, when restricted maximum likelihood (REML) is used, the deviance value shows how well only the variance estimates fit the data and the regression coefficients play no role in this computation (Peugh, 2010). So, either FIML or REML can be used to assess whether τ_{11} is nonzero, but FIML should be used if there is an interest in comparing models regarding coefficients in addition to variance components.

Referring back to our particular illustration, we implemented the nonparametric bootstrap procedure using our data and including 1,500 replications (i.e., 1,500 samples from our data with replacement). The annotated R code included in Appendix A incorporates the relevant command lines. Results indicated that the 95% bootstrap confidence interval for τ_{11} excludes zero and ranges from .004 to .046. Also, results shown in Table 1 indicate that, based on FIML, the model at Step 3 fits the data better than the model at Step 2, also suggesting a nonzero τ_{11} (i.e., deviance of 1,477.6−1,469.5 = 8.1; $p < .05$). For the sake of completeness, Table 1 also includes deviance statistics comparing the model at Step 2 with the one at Step 1 (i.e., deviance of 1,637−1,478 = 159; $p < .01$), and the model at Step 4 compared to the model at Step 3 (i.e., deviance of 1,478−1,462 = 16; $p < .01$). We also computed deviance statistics using REML. As expected, values become smaller (i.e., better fit) as we progress through the models shown in Table 1 and are as follows: 1,641, 1,492, 1,483, and 1,481.

Also as expected, the deviance statistics are overall larger (i.e., worse fit) compared to FIML because REML estimates compute fit based on differences in variance components only.

Note that each of the aforementioned tests regarding the hypothesis that τ_{11} is zero relies on null hypothesis significance testing. Thus, like all such tests of significance, statistical power is an important consideration. In other words, to be informative, such tests need to have sufficient levels of power so as to be able to detect an existing nonzero value of τ_{11} in the population. Tests regarding τ_{11} rely on degrees of freedom determined by the number of L2 units (e.g., teams), which is usually much smaller than a study's total sample size regarding lower-level units (e.g., individual employees). For example, Dalton, Aguinis, Dalton, Bosco, and Pierce's (2012) Study 1 included a review of articles published in *Journal of Applied Psychology, Personnel Psychology*, and *Academy of Management Journal* and reported median L1 sample sizes of 198, 204, and 161, respectively. In contrast, a review by Mathieu et al. (2012) including 79 multilevel investigations published in the *Journal of Applied Psychology* between 2000 and 2010 indicated that the median L2 sample size was only 51. Given that most same-level research relying on degrees of freedom based on total sample size is notoriously underpowered (Aguinis et al., 2005; Maxwell, 2004) and that multilevel modeling is usually conducted with L2 sample sizes that are much smaller, we anticipate that many tests regarding τ_{11} may also be underpowered. In other words, it is possible that in many situations there may be an incorrect conclusion that τ_{11} is not different from zero due to insufficient statistical power. As noted by an anonymous reviewer, the default position should be that if τ_{11} is not found to be different from zero, then one should not proceed with tests for possible specific cross-level interaction effects. However, to balance Type I and Type II error considerations, our recommendation is to proceed with the cross-level interaction test even when the null hypothesis of no slope variance is retained when there is a strong theory-based rationale for a particular hypothesis. Also, the fact that the null hypothesis that τ_{11} is zero was not rejected should be acknowledged explicitly so that future research can attempt to replicate the results obtained.

Using a typical χ^2 critical value with two degrees of freedom (one for τ_{11} and one for τ_{01}) to compare the models is overly conservative (i.e., likely to lead to a Type II error rate—not reject a false null hypothesis of no difference between the models). Accordingly, as a third option in terms of understanding whether τ_{11} is different from zero, Stram and Lee (1994) argued that a more appropriate distribution for such tests is a mixture of two chi-square distributions. Subsequently, Crainiceanu and Ruppert (2004) developed a method that simulates the deviance for the model with only a random intercept when testing whether the variance of slopes is significant and Scheipl, Greven, and Kuechenhoff (2008) demonstrated that the procedure is superior to competing tests (e.g., *F*-tests and tests that use critical values from a mixture of chi-square distributions) in terms of controlling Type I error rates and has similar statistical power. The procedure involves evaluating whether the variance component differs from zero by calculating the proportion of simulated deviances that exceed the sample deviance (i.e., the *p* value). Appendix A includes an R function in the RLRsim package for testing the statistical significance of variance components (Scheipl et al., 2008). Using our illustrative data, results indicated that the *p* value is .0013 and the bootstrap resampling results indicated that the 95% confidence interval for τ_{11} excludes zero and ranges from .004 to .046.

In sum, results based on our illustrative data file suggest that the relationship between LMX and individual empowerment varies depending on team membership. More precisely, results summarized in Table 1 show that the variance in slopes across groups is $\tau_{11} = .025$, and results based on the bootstrap confidence interval, the −2 log-likelihood, and the Crainiceanu and Ruppert (2004) test suggest that this value is unlikely to be zero in the population. In our example, results provide evidence in support of team-level differences in the nature of the relationship between LMX and individual empowerment which suggest the need to understand what may be the variable(s) that explain such variability. We address this issue next.

Step 4: Cross-Level Interaction Model

The fourth and final step in the model building process involves understanding whether a particular L2 variable is able to explain at least part of the variance in slopes across teams. Referring back to our substantive illustration, we would like to know whether leadership climate moderates the relationship between LMX and empowerment across teams. To do so, we begin building the *cross-level interaction model* with Equation 13 (identical to Equation 1):

$$\text{Cross-Level Interaction Model (Level 1): } Y_{ij} = \beta_{0j} + \beta_{1j}\left(X_{ij} - \overline{X}_j\right) + r_{ij} \qquad (13)$$

Then, we allow both intercepts and slopes to vary across teams as follows:

$$\text{Cross-Level Interaction Model (Level 2): } \beta_{0j} = \gamma_{00} + \gamma_{01}\left(W_j - \overline{W}\right) + u_{0j} \qquad (14)$$

$$\text{Cross-Level Interaction Model (Level 2): } \beta_{1j} = \gamma_{10} + \gamma_{11}\left(W_j - \overline{W}\right) + u_{1j} \qquad (15)$$

The difference between Equation 15 (cross-level interaction model) and Equation 11 (RIRSM) is that Equation 15 includes the L2 predictor hypothesized to play a moderating role. We are no longer solely interested in whether there is variance in slopes across teams—that was the purpose of the previous step. Now, we are interested in understanding whether such variance can be explained by a particular L2 predictor (i.e., leadership climate).

In Equation 15, the moderating effect of leadership climate on the relationship between LMX and empowerment is captured by γ_{11}. Equivalently, γ_{11} is the cross-level interaction of LMX and leadership climate on empowerment. That is, γ_{11} represents the change in the slope of empowerment on LMX scores across teams when leadership climate increases by 1 point. For example, a result that γ_{11} is positive indicates that LMX is more strongly related to empowerment in teams with more positive leadership climate compared to teams with less positive leadership climate.

Substituting the L2 Equations 14 and 15 into the L1 Equation 13 leads to a combined model as follows:

Cross-Level Interaction Model (Combined):

$$Y_{ij} = \gamma_{00} + \gamma_{01}\left(W_j - \overline{W}\right) + \gamma_{10}\left(X_{ij} - \overline{X}_j\right) + \gamma_{11}\left(X_{ij} - \overline{X}_j\right)\left(W_j - \overline{W}\right) + u_{0j} + u_{1j}\left(X_{ij} - \overline{X}_j\right) + r_{ij} \qquad (16)$$

Equation 16 resembles the more familiar MMR model, which also includes the constituent linear terms. However, in contrast to the MMR model, Equation 16 includes the terms involving u_{0j} and u_{1j}, which vary across L2 units, and, as mentioned earlier, this is why they are labeled *random effects*. On the other hand, γ_{00}, γ_{01}, γ_{10}, and γ_{11} are constant across L2 units, so they are labeled *fixed effects*.

Results using our illustrative data provide evidence in support of the cross-level interaction effect we tested. Table 1 shows that the slope of individual empowerment on LMX is expected to equal $\gamma_{10} = 0.269$ for teams with an average leadership climate. However, the relationship between individual LMX and individual empowerment becomes stronger, by $\gamma_{11} = 0.104$ units, as a team's leadership climate increases by one unit.

Finally, an issue to consider is the possibility that as a result of implementing Step 2, results may suggest a nonsignificant L1 relationship between X and Y (i.e., $\gamma_{10} = 0$). In such instances, researchers may be hesitant to proceed with Step 3 and investigate possible cross-level interactions. However, there could be variability in group slopes although $\gamma_{10} = 0$. Accordingly, if there is a theory-based rationale for examining cross-level interaction effects, we recommend proceeding with Step 3 regardless of the statistical significance of the direct effect for X. Moreover, standard practice when estimating interactions is to include lower-level effects, regardless of statistical significance, and this is a correct practice for the following reason. Consider the L2 equation for slopes for Step 3 (see Equation 11 above). In Equation 11, a nonsignificant relationship between X and Y implies that $\beta_{1j} = u_{1j}$ (i.e., on average the relationship is zero, but groups deviate from zero by u_{1j}). Now, consider Equation 15, where the cross-level interaction effect is estimated by including $W_j - \overline{W}$. In Equation 15, γ_{11} is the cross-level interaction effect and γ_{10} is the relationship between X and Y for groups with $W_j - \overline{W}$. It is possible that $\gamma_{10} = 0$ for Model 3 in Equation 11, but $\gamma_{10} \neq 0$ after W_j is included in the equation and centered by the mean or some other value. Consequently, leaving X out of Step 4 will force $\gamma_{10} = 0$ at the point where $W_j - \overline{W}$. Accordingly, we recommend including X in the model to account for the fact that the relationship between X and Y may not be zero for the value at which W_j is centered.

Multilevel Modeling Assumptions

Although not specific to tests of cross-level interaction effects, there are several assumptions that underlie multilevel modeling in general, which parallel the usual OLS regression assumptions. Violating these assumptions can have consequences in terms of the validity of the inferences made from the results. Specifically, violating assumptions can lead to model misspecification—a misrepresentation of relationships among variables. Thus, it is important to assess compliance with these assumptions by using methods described by Raudenbush and Bryk (2002) and Snijders and Bosker (2012).

First, function forms are assumed to be correctly specified at each level (e.g., a linear, quadratic, or higher-order polynomial). For example, if there is a pattern of curvilinearity such that very high levels of LMX scores are associated with a decrease in empowerment scores (i.e., inverted U-shaped form), not including quadratic terms in the equation would preclude the identification of such nonlinear relationship and, likely, result in interventions with detrimental consequences for individuals and organizations (Pierce & Aguinis, 2013).

Second, there are several assumptions regarding residuals: (a) L1 residuals (i.e., r_{ij}) are assumed to be normally distributed and have a mean of zero, (b) L2 residuals (i.e., u_{0j} and u_{1j}) are assumed to conform to a multivariate normal distribution and also have means of zero, (c) L1 residual variance is assumed to be constant (i.e., homoscedasticity) both within and between L2 units, and (d) L1 residuals and L2 residuals are assumed to be uncorrelated. In addition to the overall concern about model misspecification, violating residual-based assumptions can lead to invalid hypothesis tests because standard errors may be grossly misspecified (Snijders & Bosker, 2012); so, there are alternatives that allow researchers to relax some assumptions regarding L1 and L2 residuals and prevent such errors (e.g., Culpepper, 2010). Finally, violating the assumption that L1 residuals and L2 residuals are uncorrelated implies the possibility of crossover relationships. For example, individuals could be influenced by more than one leader (i.e., crossover leadership influences), resulting in a "cross-classified" situation. Such crossover influences may be more common than presently acknowledged (Han, 2005; Mathieu & Chen, 2011). Accordingly, Han (2005) proposed methods for considering crossover influences and modeling them. The presence of such crossover influences may not only affect the accuracy of tests of cross-level interaction hypotheses, but lead to overall model misspecification in general.

Best-Practice Recommendations

Although there is increasing awareness regarding the need for multilevel modeling in management research, there are important questions about what researchers should do prior and after data collection to improve the accuracy of substantive conclusions regarding cross-level interaction effects. We compiled a list of the most persistent and challenging questions by conducting a systematic review using the entire archives of two listservs: RMNET (Research Methods Division of the Academy of Management) and MULTILEVEL (list specifically devoted to multilevel analysis). At the time of this writing, RMNET includes approximately 1,000 members and MULTILEVEL includes more than 1,400 members. The goal of our review was to gather all the questions posted on these listservs that are directly or indirectly related to the estimation and interpretation of cross-level interaction effects in multilevel modeling. In other words, our review provided us with information on the most frequent and challenging issues faced by researchers in their attempts to test hypotheses about cross-level interaction effects.

We conducted our search using the terms *cross-level, interaction, multilevel, moderator, moderate*, and *moderating*. Next, we discuss issues for which there is sufficient evidence to support a particular best-practice recommendation. We offer recommendations for actions researchers can take prior to data collection and recommendations researchers can implement after data have been collected. Table 2 offers a list of these recommendations, which we describe in more detail next.

Pre–Data Collection Recommendations

Issue 1: What is the operational definition of a cross-level interaction effect? A frequently asked question regarding cross-level interaction effects refers to the very definition of this effect.

Table 2
Summary of Best-Practice Recommendations for Estimating Cross-Level Interaction Effects Using Multilevel Modeling

Pre–Data Collection Recommendations

- *Recommendation 1: Defining the Cross-Level Interaction Effect.* Clearly and unambiguously identify and define the cross-level interaction effect. In the combined Level 1 and Level 2 equation, the cross-level interaction effect is the coefficient associated with the product term between the Level 1 and Level 2 predictors.
- *Recommendation 2: Calculating Statistical Power.* Design the multilevel study so that it will have sufficient statistical power to detect an existing cross-level interaction effect. Use the R code provided by Mathieu et al. (2012) and available at http://mypage.iu.edu/~haguinis to understand trade-offs in research design and measurement choices and allocate resources accordingly. Compute power after data collection if the cross-level interaction effect is not found. If power was sufficient, then one can have confidence that the effect does not exist in the population; if power was insufficient, then report the power value obtained and report that results are inconclusive because of the possibility that the population effect exists but was not detected in the sample.
- *Recommendation 3: Testing Hypotheses About Different Types of Moderator Variables.* Plan to test hypotheses about cross-level interaction effects involving Level 2 continuous or categorical variables, but be aware of resulting differences in how to interpret the observed effect.

Post–Data Collection Recommendations

- *Recommendation 4: Rescaling (i.e., Centering) Predictor Variables.* In most cases, center the Level 1 predictor by team mean scores (i.e., group-mean centering) to improve the interpretation of the cross-level interaction effect. However, theory-based considerations should dictate the chosen approach to rescaling.
- *Recommendation 5: Graphing the Cross-Level Interaction.* Graph the cross-level interaction effect to understand its nature and direction. However, do not use the graph to draw conclusions about the size or importance of the effect.
- *Recommendation 6: Interpreting Cross-Level Interaction Effects.* Interpret the Level 1 predictor or the Level 2 predictor as the moderator based on substantive and conceptual interests because the cross-level interaction effect is symmetrical in nature. In most cases, the Level 2 (or higher-level) predictor will serve the role of the moderator variable.
- *Recommendation 7: Estimating Multiple Cross-Level Interaction Effects.* Include all cross-level interaction effects as part of the same model when testing more than one cross-level interaction effect. However, strong theory-based considerations as well as other situations (e.g., models may not converge, they may crash, or run out of degrees of freedom) may justify conducting a separate test with each interaction effect.
- *Recommendation 8: Testing Cross-Level Interactions Involving Three or More Levels of Analysis.* Conduct tests of three-way and higher-order cross-level interaction effects following the same procedures as those for two-way interactions, but be mindful that adequate sample sizes will be required for each of the levels involved.
- *Recommendation 9: Assessing Practical Significance.* Compute the size of the cross-level interaction effect based on its predictive power as well as its explanatory power and place resulting effect sizes within context to understand their importance for theory and practice.
- *Recommendation 10: Reporting Results.* Report complete results regarding each of the steps of the multilevel model building process including all coefficients and their standard errors as well as variance components—see Table 1 for a template.

First, there is a question of whether τ_{11} (i.e., the variance of slopes across groups) is the cross-level interaction effect. Second, there is a question of whether γ_{11} in the hierarchical linear model in Equation 15 can truly be called an interaction effect given that it is associated with a term that does not involve a product between two variables, but with one variable only (i.e., W).

First, let's consider τ_{11}. As mentioned earlier, a nonzero τ_{11} means that the slope of the L1 criterion on the L1 predictor varies across higher-level units (e.g., teams). Referring back to our substantive example, a nonzero τ_{11} means that the effect of LMX on empowerment is not homogeneous across teams. However, this heterogeneity may be due to one or more potential L2 moderators. In our particular example, we tested the potential moderating effect of leadership climate. However, we could have considered additional moderators, instead or in addition to, leadership climate. Thus, similar to the assessment of moderating effects in the context of meta-analysis, the presence of heterogeneity is an indication that the search for particular moderators is warranted, but this is not the interaction effect per se (Aguinis, Gottfredson, & Wright, 2011).

Now, let's consider the meaning of γ_{11} in the context of two extreme situations. First, assume that γ_{11} in Equation 15 is zero. This would mean that L2 variable *W* does not explain variance in the slope of the L1 *Y* outcome on the L1 *X* predictor across teams. Thus, for every unit increase in the higher-order variable (*W*), the relationship (slope) between L1 *X* and L1 *Y* remains unchanged. Now, assume that γ_{11} in Equation 15 is a very large and positive number, which implies that a small change in the value of *W* is associated with a large change in the slope of L1 outcome *Y* on L1 predictor *X* across teams. Thus, a nonzero γ_{11} means that the L1 effect of *X* on *Y* is distributed across L2 units—and this is the reason why Raudenbush and Bryk (2002) used the term *distributive effect* to refer to the cross-level interaction effect.

The fact that γ_{11} represents the cross-level interaction effect, also labeled the moderating effect of *W* on the *X*-*Y* relationship, is perhaps seen more easily by referring to the combined Equation 16. In Equation 16, which has a familiar form that closely matches that of MMR, γ_{11} is associated with the product term between rescaled L1 predictor *X* and rescaled L2 predictor *W*. In contrast, in the L2 Equation 15, γ_{11} is a coefficient predicting slope values, which is a model not as familiar to management researchers compared to a model that predicts criterion *Y* scores.

In sum, the coefficient γ_{11} is interpreted as the cross-level interaction effect regardless of whether it is obtained by using Equation 15 (i.e., predicting β_{1j} based on $W_j - \overline{W}$) or Equation 16 (i.e., predicting γ_{ij} based on $(X_{ij} - \overline{X}_j)(W_j - \overline{W})$). The variance of slopes across groups τ_{11} is not the cross-level interaction effect because, although it provides information on the extent to which the slope of the L1 criterion on the L1 predictor varies across higher-level units (e.g., teams), it does not provide information on the particular variable(s) that are associated with this variability. The specific interpretation regarding the meaning of the cross-level interaction effect γ_{11} will depend on the approach adopted regarding rescaling, which is in turn dictated by theory-based considerations. We describe rescaling in more detail in our discussion of Issue 4.

Issue 2: What is the statistical power to detect an existing cross-level interaction effect?. A second pre–data collection question that has appeared frequently refers to research design and statistical power. Specifically, researchers are interested in understanding how large a sample size should be to detect existing cross-level interaction effects. Statistical power is a complex issue in the context of cross-level interaction effects and tools such as *Optimal Design* (Raudenbush, 1997; Spybrook, Raudenbush, Congdon, & Martinez, 2009) and *Power*

IN Two-level designs (PINT; Bosker, Snijders, & Guldemond, 2003) do not provide power estimates for cross-level interaction tests. In fact, Snijders (2005: 1572) noted that "for the more general cases, where there are several correlated explanatory variables, some of them having random slopes, such clear formulae are not available." Accordingly, Scherbaum and Ferreter (2009: 363) concluded that "estimates of statistical power of cross-level interactions are much more complex than the computations for simple main effects or variance components . . . and there is little guidance that can be provided in terms of simple formulas."

Because of a lack of analytic solutions, Mathieu et al. (2012) conducted a Monte Carlo simulation to understand the impact of various factors that affect the power of cross-level interaction tests. Results of their study revealed that the power to detect cross-level interactions is determined primarily by the magnitude of the cross-level interaction effect, the variance of L1 slopes across L2 units, and by L1 and L2 sample sizes. Researchers usually do not have control over the size of the cross-level interaction effect or the variance of L1 slopes across L2 units. On the other hand, although there may be practical and resources-related constraints, researchers may be able to increase L1 and L2 sample sizes to increase power. As concluded by Raudenbush and Liu (2000), L1 sample size is most relevant for the statistical power to detect L1 direct effects and L2 sample size is most relevant for the statistical power to detect L2 direct effects. Thus, researchers interested in both types of direct effects face a difficult dilemma in terms of the allocation of the research budget, which is typically limited and may not allow for the allocation of resources to increase both L1 and L2 sample sizes.

For the particular case of power to detect cross-level interaction effects, Mathieu et al.'s results revealed that the average L1 sample size has a relative premium of about 3:2 as compared to the L2 sample size. Moreover, Mathieu et al.'s results indicated that

> both levels' sample sizes interacted significantly with the magnitude of the cross-level interaction, and with the variability of the Level 1 slopes. . . . Ultimately, the decision as to focus on maximizing Level 1 versus Level 2 sample sizes may come down to what other parameters are of interest in an investigation. . . . [I]f besides the cross-level interaction a researcher is interested in testing a lower-level direct effect, then perhaps Level 1 sample sizes are most important. Alternatively, if the researcher is also interested in testing cross-level direct effects, that may suggest emphasizing the number of units that are sampled. (Mathieu et al., 2012: 960)

In addition to the simulation, Mathieu et al. (2012) conducted a power analysis based on articles published in *Journal of Applied Psychology* from 2000 to 2010 and found that power has been quite low. Specifically, at the $\alpha = .05$ level, the average power was .40, and at $\alpha = .01$, the average statistical power value was only .22. Thus, statistical power to detect cross-level interactions is substantially below the conventional .80 level, and researchers interested in testing interaction effects in the context of multilevel modeling should indeed be concerned about statistical power. In other words, given an existing population cross-level interaction effect, the typical probability of actually detecting the effect is less than the flip of a coin. Although there is no way to know whether a particular population effect exists, low statistical power means that it is likely that many researchers have erroneously concluded that a cross-level interaction effect is not different from zero.

Based on these results, Mathieu et al. (2012) created a computer program available online at http://mypage.iu.edu/~haguinis/crosslevel.html that allows researchers to estimate the power of their cross-level interaction test prior to data collection. The program can be used to gather important information in terms of solving a possible dilemma regarding the decision to increase the number of L1 compared to L2 units. For example, a researcher can use the program under two different scenarios. Hypothetical Scenario A would involve the possibility of increasing the number of individuals per team by 5. Hypothetical Scenario B would involve holding the number of individuals per team constant but, instead, increasing the number of teams from 50 to 80. Using these different values as input in the power calculator allows for an understanding of the statistical power associated with each of these scenarios, and results can be used for design planning and as a guide in making more informed and better decisions about how to allocate research resources prior to data collection. In addition, the power calculator can also be used to make better-informed decisions about substantive conclusions. Specifically, if a cross-level interaction effect hypothesis is not supported and the power calculator suggests that power was sufficient, then one can have confidence that the effect does not exist in the population. On the other hand, if power was insufficient, then researchers need to report the power value obtained and, unfortunately, report that results are inconclusive because of the possibility that the population effect exists but was not detected in the sample.

In sum, given the possible trade-offs between L1 and L2 sample sizes as well as interactive effects of the various factors on power (e.g., size of the cross-level interaction effect, variance of L1 slopes across L2 units), our recommendation is to abandon popular rules of thumb such as the “30-30 rule” (i.e., having at least 30 upper-level units with at least 30 lower-level entities in each; Kreft & De Leeuw, 1998). Also, researchers should not assume that a particular sample size is sufficient to detect an existing effect—for example, Liden and Antonakis (2009: 1599) asserted that “30-50 [i.e., at least 30 upper-level units and at least 50 lower-level entities in each] . . . should be sufficient to estimate multilevel models correctly.” Instead, researchers should use the Mathieu et al. (2012) power calculator a priori to make decisions about research design features and also post hoc to understand whether published studies reached sufficient levels of statistical power to detect existing effects.

As an illustration, we used Mathieu et al.’s (2012) power calculator with our own data set. Necessary input includes L1 and L2 sample sizes, ICC, and several of the parameter estimates in Equation 16. As noted earlier, the program can be used a priori with various sample sizes to understand what particular combination of L1 and L2 sample size would lead to a desired power level (e.g., .80). Alternatively, the program can also be used after a study is conducted to understand the probability that the particular L1 and L2 sample sizes used allowed for a detection of an existing cross-level interaction effect of a particular size.

Appendix B includes the annotated R code we used for our power calculation (this code is also available at http://mypage.iu.edu/~haguinis). We used Chen et al.’s (2007) estimates to guide us on reasonable values for the ICC of X_{ij} (i.e., .12), intercept variance between teams (roughly .1), and within-team variance in individual empowerment (approximately .8). Last, we also needed estimates for the grand mean relationship between X_{ij} and Y_{ij}, the variability of slopes, and the magnitude of the cross-level interaction effect. We chose modest values (compared to typical values found in published articles and reported by Mathieu et al., 2012) for the relationship between X_{ij} and Y_{ij} (.4) and the standard deviation of slopes

(.1). We may not have any indication as to the expected size of the cross-level interaction effect, so we could use the calculator with a moderate effect (also, in relationship to values reported by Mathieu et al., 2012). We emphasize that the choice for a particular targeted effect size should be guided by theoretical (i.e., What does past research show regarding the size of the effect?) as well as practical (i.e., What is a sufficiently large effect worth detecting?) considerations. However, given the pedagogical and illustrative nature of our power analysis, we simplified the process of selecting our targeted effect size. Entering these values into the power calculator provides evidence that the power to detect a moderate effect for our proposed study design was .82.

Issue 3: Is it possible to test for cross-level interaction effects involving a categorical L2 or standardized predictor? A third issue that has been frequently raised refers to the possibility that the multilevel model may include a L2 variable that is not continuous in nature but, rather, has discrete categories (e.g., old vs. new compensation system). In other words, the question is whether a researcher can test a hypothesis about cross-level interaction effects involving a L2 predictor that is categorical. Fortunately, this is possible. However, similar to the use of categorical predictors in the context of MMR (Aguinis, 2004), the interpretation of the cross-level interaction effect is in reference to differences in relationships between two or more groups. Based on our discussion so far, and specifically referring to Equation 16, γ_{11} is interpreted as a change in the slope of Y on X across teams associated with a 1-unit change in W. Now, assume that W is a binary variable that was coded as 1 = new compensation system and 0 = old compensation system. The hypothesis is that the relationship between LMX scores and empowerment across teams will vary as a function of compensation system. If $\gamma_{11} = 1.5$, its interpretation is that the slope of Y on X is 1.5 points larger for teams in the new compensation system (i.e., coded as 1) compared to teams in the old compensation system (i.e., coded as 0). In other words, there is a stronger relationship between LMX and empowerment for individuals working in teams under the new compensation system. If the binary moderator is group-mean centered, the mean is a proportion of the category scored 1, but the interpretation is similar in the sense that the coefficient refers to changes in the slope of Y on X for teams coded as 1 compared to teams coded as 0.

When the categorical L2 predictor includes more than $k = 2$ values, it is necessary to create $k - 1$ dummy codes, which are added to Equation 16. The process is similar to creating dummy codes in the context of MMR (see Aguinis, 2004, chap. 8). Assuming a L2 predictor with $k = 3$, the two dummy codes $W_{j(1)}$ (e.g., comparison of category 1 vs. 2) and $W_{j(2)}$ (e.g., comparison of category 1 vs. 3) are included in Equation 16 as follows:

$$\begin{aligned} Y_{ij} &= \gamma_{00} + \gamma_{01(1)}\left(W_{(1)j} - \overline{W}_{(1)}\right) + \gamma_{01(2)}\left(W_{(2)j} - \overline{W}_{(2)}\right) + \gamma_{10}\left(X_{ij} - \overline{X}_j\right) \\ &\quad + \gamma_{11(1)}\left(X_{ij} - \overline{X}_j\right)\left(W_{j(1)} - \overline{W}_{(1)}\right) + \gamma_{11(2)}\left(X_{ij} - \overline{X}_j\right)\left(W_{j(2)} - \overline{W}_{(2)}\right) \\ &\quad + u_{0j} + u_{1j}\left(X_{ij} - X_j\right) + r_{ij} \end{aligned} \tag{17}$$

Note that, similar to the situation involving two categories only, the interpretation of the cross-level interaction effect coefficients must consider which category was coded as 1 and 0 for each dummy variable. So, for example, assuming that the categories are three locations: (a) Colorado, (b) Indiana, and (c) Texas, and that $W_{j(1)}$ involves a comparison of Colorado (coded as 1) and Texas (coded as 0). A statistically significant $\gamma_{11(1)} = 2$ would mean that the slope of empowerment on LMX is 2 points larger for teams in Colorado compared to teams in Texas.

Finally, some researchers choose to standardize predictors (i.e., rescale raw scores so they have a mean of zero and standard deviation of one) to be able to interpret results referring to *SD* units instead of the units used in the original scales (e.g., 7-point Likert-type scales). In fact, this is precisely what Chen et al. (2007) did in their study. In such situations, referring back to Equation 16, γ_{11} is the expected change in the size of the slope of LMX on empowerment in *SD* units that is associated with a 1-*SD* unit increase in the L2 predictor (i.e., leadership climate).

Post–Data Collection Recommendations

Issue 4: How should I rescale (i.e., center) predictors and why? As noted earlier, rescaling predictors is common when conducting multilevel analyses to help in the interpretation of results (Dalal & Zickar, 2012). The two main rescaling approaches are group-mean centering (which we used in our article) and grand-mean centering (Enders & Tofighi, 2007). A third option is to not rescale predictors at all, but this is not recommended because in many situations the resulting parameter estimates will be uninterpretable. Specifically, if we use raw (i.e., uncentered) scores instead of rescaled scores, β_{0j} in Equation 1 represents the predicted level of empowerment for a LMX score of zero. But, this would be a meaningless interpretation because the LMX scale ranges from 1 to 7 and does not include zero as a possible value. Moreover, a LMX score of zero may not be meaningful at the construct level because it is possible to have a low or high LMX, but it is not possible to have no LMX at all. Most organizational behavior and human resource management measures do not have a true zero value because they are not ratio in nature. Similarly, many measures often used in entrepreneurship research (e.g., entrepreneurship orientation; Covin & Wales, 2012) and strategy (e.g., firm resources and capabilities; Barney, Ketchen, & Wright, 2011) also do not have a true zero value. Thus, rescaling is needed in most studies in these domains. Alternatively, some financial measures used in strategy do have a meaningful zero point (e.g., return on assets, return on investment; Dalton & Aguinis, 2013), so, in these cases, rescaling may not be needed.

Group-mean centering changes the mean and correlation structure of the data, causing the L1 predictors to be uncorrelated with the L2 predictors (Enders & Tofighi, 2007). Also, in the first section of our article we used group-mean centering for the L1 predictor scores to interpret the resulting coefficients in reference to team-level average LMX scores. Alternatively, grand-mean centering involves using the mean of all scores at a particular level. So, going back to our example, grand-mean centering the L1 predictor would involve using the mean LMX scores across all 630 individuals and grand-mean centering the L2 predictor would involve using the mean team leadership score across all 105 teams. An important concern regarding the use of grand-mean centering for the L1 predictor is that γ_{11} (i.e., cross-level interaction effect coefficient) conflates the between-team and within-team effects. In other words, using grand-mean centering for the L1 predictor leads to a cross-level interaction effect coefficient that is a "mixed bag" of two separate effects: (a) a true cross-level interaction involving the upper-level moderator and the within-group variance of the lower-level predictor (which is what we are interested in estimating) and (b) a between-level interaction (i.e., an interaction between the upper-level moderator and the between-group variance of the lower-level predictor). Thus, Enders and Tofighi (2007) argued that if a researcher uses grand-mean centering for the L1 predictor, it is not possible to make an accurate, or even meaningful, interpretation of the cross-level interaction effect. Accordingly, Hofmann and Gavin (1998) concluded that group-mean centering leads to the most accurate estimates of within-group slopes and minimizes the possibility of finding spurious cross-level interaction effects. Similarly, Preacher et al. (2010) used the label "unconflated model" in referring to a model based on group-mean centered L1 predictors.

Although group-mean centering has been recommended, overall, as the best strategy in the context of testing cross-level interaction hypotheses, it is important to recognize that such a choice needs to reflect theoretical processes related to deviations from a group average (e.g., social comparison effects in team research). Moreover, Bliese (2002: 433) noted that "spurious cross-level interactions are rare, so one can generally use . . . grand-mean-centered variables to test for cross-level interactions as long as one runs an additional model with group-mean-centered variables to check for spurious interaction effects." The alternative grand mean centering strategy would involve controlling for between-group variance by estimating the interaction between the L2 predictor and the group averages for the L1 constituent linear terms. One advantage of group-mean centering is that, because there is no need to control for across-group variance (i.e., group-mean centering addressing this issue), the resulting model includes fewer parameter estimates. However, estimating cross-level interactions using group-mean centering has a different substantive interpretation than estimating interactions using grand-mean centering. As noted by an anonymous reviewer, using group-mean centering suggests that testing interactions needs to reflect theoretical processes addressing deviations from a group average such as in frog pond/social comparison effects in studies of teams. However, not all theories specifically refer to deviations from group averages or have reached that level of sophistication. In some situations, it may be more appropriate to use grand-mean centering with across-group variance controlled because a theory may address raw differences between L1 entities, not differences relative to a group average.

In short, group-mean centering the L1 predictor is the recommended approach in most situations when there is an interest in testing hypotheses about cross-level interaction effects. However, in some situations it may be more appropriate to use grand-mean centering with across-group variance controlled because a particular conceptualization may address raw differences between L1 entities rather than differences relative to a group average. Thus, the choice for a rescaling approach must be accompanied by a theory-based justification regarding the underlying process that is being modeled.

Issue 5: How can I graph a cross-level interaction effect? Another question often posted on the listservs refers to how to produce graphs to illustrate a cross-level interaction effect. Similar to the single-level context, graphs can be used to illustrate the nature of the interaction effect, but should not be used to draw conclusions about the size or importance of the effect (Aguinis, 2004). Considering the combined model in Equation 16, the expected value of γ_{ij} conditioned on values of X_{ij} and W_j can be written as:

$$E\left(Y_{ij} \mid X_{ij}, W_j\right) = \gamma_{00} + \gamma_{01}\left(W_j - \overline{W}\right) + \gamma_{10}\left(X_{ij} - \overline{X}_j\right) + \gamma_{11}\left(X_{ij} - \overline{X}_j\right)\left(W_j - \overline{W}\right) \tag{18}$$

$$= \gamma_{00} + \gamma_{01}\left(W_j - \overline{W}\right) + \left(\gamma_{10} + \gamma_{11}\left(W_j - \overline{W}\right)\right)\left(X_{ij} - \overline{X}_j\right) \tag{19}$$

In Equation 19, the relationship between X_{ij} and Y_{ij} is represented by the term preceding, $X_{ij} - \overline{X}_j$, namely $\gamma_{10} + \gamma_{11}(W_j - \overline{W})$. If W_j is a continuous variable, Equation 19 can be used to plot the X-Y relationship for any value for W_j. The equation describing the relationship between X and Y for a specific value of W_j is called a *simple regression equation* and the slope of Y on X at a single value of W_j is called a *simple slope*. Preacher, Curran, and Bauer (2006) provided specific examples as well as a description of computer programs in SAS, SPSS, and R that allow for the creation of plots to more easily understand the nature of the interaction effect, including the plotting of simple slopes. The Preacher et al. programs also allow for plotting *regions of significance*, which are values of W between which the simple slope of Y on X is statistically significant.

Equation 19 can also be used to plot the interaction effect in cases where W_j is a binary L2 variable that takes on the values of 0 and 1. Also, if W_j is a binary variable, it is easier to interpret the model coefficients if W_j is not rescaled. Consequently, the predicted values of Y_{ij} for the two groups defined by W_j are

$$E\left(Y_{ij} \mid X_{ij}, W_j = 0\right) = \gamma_{00} + \gamma_{10}\left(X_{ij} - \overline{X}_j\right) \tag{20}$$

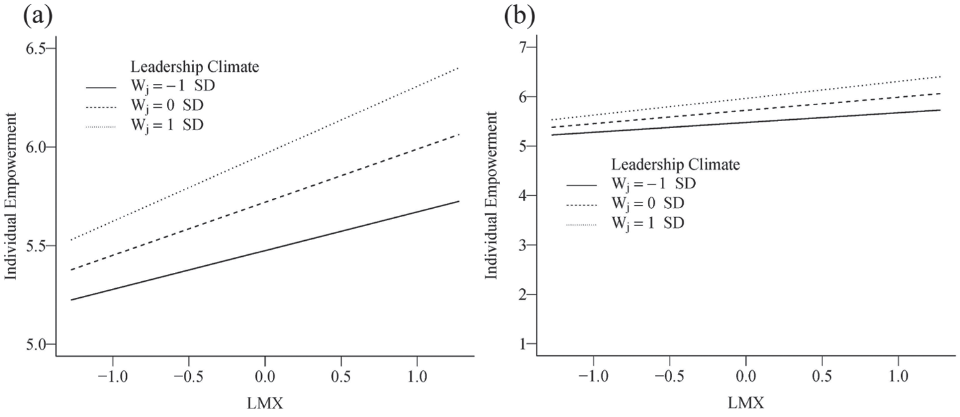
$$E\left(Y_{ij} \mid X_{ij}, W_j = 1\right) = \gamma_{00} + \gamma_{01} + \left(\gamma_{10} + \gamma_{11}\right)\left(X_{ij} - \overline{X}_j\right) \tag{21}$$

Now, the X-Y relationship can be plotted for each group. To do so, we can use values of one standard deviation below the mean, the mean, and one standard deviation above the mean for X. Using these particular values is recommended because they allow for an understanding of the nature of the relationship across a wide range of X scores (Aiken & West, 1991). Moreover, it may also be useful to choose additional values that may be informative in specific contexts.

The plots included in Figure 3 show the cross-level interaction effect in our illustrative data. One important issue to consider concerns the axis for the Y scale. In many published articles, researchers use reduced scales for the Y axis (e.g., 4 scale points instead of 7). Such reduction in the scale gives the false impression that the effect is more important because the slopes seem steeper and also more different from each other. So, it is acceptable to reduce the length of the axis to understand the nature of the interaction, as we have done in Figure 3's Panel (a) (Panel (b) includes the same plot with the full Y scale represented along the axis). But it is not acceptable to do so and then make statements about how "important" an effect is given the degree of steepness of the slope on the graph. The annotated R code in Appendix A also includes the necessary commands to create the plots included in Figure 3.

Issue 6: Are cross-level interaction effects symmetrical? Another issue related to the interpretation of cross-level interaction effects refers to whether such effects are symmetrical.

Figure 3
Plots of Moderating Effect of L2 Variable Leadership Climate on the Relationship Between L1 Variable Quality of Leader-Member Exchange (LMX) and L1 Variable Individual Empowerment



Note. Panel a: reduced y-axis scale; Panel b: entire y-axis scale

In other words, can we say that the L2 variable moderates the effect of the L1 predictor on the L1 criterion and, also, that the L1 predictor moderates the effect of the L2 predictor on the L1 criterion? From a statistical standpoint, it is just as appropriate to label the L2 predictor as the moderator of the effect of the L1 predictor *X* on the L1 outcome *Y* as it is to label the L1 predictor *X* as the moderator of the effect of the L2 predictor *W* on the L1 outcome *Y* because cross-level interactions are symmetrical. In multilevel modeling, it is usually the case that L2 is labeled as the moderator because, conceptually, it seems more appropriate to frame the higher-level variable as the contextual factor that affects the relationship between lower-level variables. Specifically, in our discussion regarding how to graph cross-level interaction effects, we used the L2 variable as the moderator. However, referring back to Equation 16, the value for γ_{11} is obviously the same whether it is associated with $(X_{ij} - \overline{X}_j)(W_j - \overline{W})$ or with $(W_j - \overline{W})(X_{ij} - \overline{X}_j)$. Thus, the choice to interpret *W* or *X* as the moderator is based on conceptual reasons.

Referring back to our substantive illustration, we could have chosen to state the cross-level interaction effect hypothesis using either of the following forms:

Hypothesis 1a—L2 moderator: The effect of individual LMX on individual empowerment will be moderated by leadership climate such that higher levels of leadership climate will lead to a stronger LMX–empowerment relationship compared to lower values of leadership climate.

Hypothesis 1b—L1 moderator: The effect of leadership climate on individual empowerment will be moderated by individual LMX such that higher levels of LMX will lead to a stronger leadership climate–empowerment relationship compared to lower values of LMX.

In short, a substantive interest in studying either the L2 or the L1 predictor as a moderator dictates how a researcher will conceptualize the nature of the cross-level interaction effect, but L2 is usually labeled as the moderator due to conceptual reasons. Choices about how to phrase the cross-level interaction effect hypothesis as well as graph the resulting effect will follow directly from the choice of which variable is labeled the moderator.

Issue 7: How can I estimate more than one cross-level interaction effect? Also related to post–data collection issues are questions about more complex models. For example, assume we are interested in testing two cross-level interactions: (a) interaction between L2 predictor *W* and L1 predictor *X* on *Y* and (b) interaction between L2 predictor *Z* and L1 predictor *X* on *Y*. A frequent question is whether testing these hypotheses should be done sequentially in separate models or simultaneously in one single model including both $\left(X_{ij}-\overline{X}_j\right)\left(W_j-\overline{W}\right)$ and $\left(X_{ij}-\overline{X}_j\right)\left(Z_j-\overline{Z}\right)$.

Overall, the recommendation is to test both interaction effects as part of one combined model so that each estimated effect is adjusted for all the theoretically relevant components. If the hypothesized cross-level interaction effects are evaluated in separate models, it is possible that these effects will be upwardly biased due to possible nonzero intercorrelations between the various interaction effects. However, given that most cross-level interaction tests are likely to be insufficient regarding statistical power (see Issue 2), a strong theory-based rationale for the presence of such effects may justify conducting the tests separately. From the perspective of a trade-off between Type I (i.e., false positive) and Type II (i.e., false negative) statistical errors, this approach would be equivalent to conducting follow-up comparisons in ANOVA without first conducting an omnibus test.

An additional consideration in implementing our recommendation to test both interaction effects as part of one combined model is that complex models may not converge, they may crash, or run out of degrees of freedom. In such situations, and absent a strong theory-based rationale for testing models separately, our recommendation is to proceed with testing models separately but then report results in a transparent and open manner (Aytug, Rothstein, Zhou, & Kern, 2012). In other words, it is necessary to be clear about this limitation (i.e., models were tested separately), the reason why (e.g., the combined model crashed), and consequences of the limitation (i.e., the need to replicate results in future research due to a possible inflation of Type I error rates; Brutus, Aguinis, & Wassmer, 2013).

Finally, an issue related to complex models in general is that they may not converge. There are several reasons that may lead to this situation. For example, a model may not converge when certain algorithms are used (Wolfinger & O'Connell, 2007), the random effects are highly correlated, or the model is misspecified (e.g., the model may be too complex for the data). There are several possible courses of action when models do not converge. An initial course of action is to use a different software program. If the model still does not converge, a second alternative is to center predictors because centering can help reduce correlations among random intercepts and slopes (Gelfand, Sahu, & Carlin, 1995). Ultimately, the source of the problem may be that the model is misspecified or too complex for the data in hand. In such cases, the only solution may be to simplify the random effects structure of the model.

Issue 8: How can I estimate cross-level interaction effects involving variables at three different levels of analysis? Another issue regarding more complex models involves the possibility of testing cross-level interactions involving more than two levels of nesting. As one illustration, a researcher may be interested in testing a three-level interaction effect of LMX (Level 1, *X*), leadership climate (Level 2, *W*), and organizational culture (Level 3, *Q*) on individual empowerment. In such three-level models, team-level relationships are allowed to vary across higher-level units (e.g., organizations, geographic regions). Other situations that may involve three levels of analysis are studies that rely on experience sampling methodology or other types of diaries (e.g., Uy, Foo, & Aguinis, 2010). In such situations, there are observations of individuals over time (i.e., observations are nested within individuals) and individuals are nested within units (e.g., teams). So, in this situation there are three levels of analysis with two levels of nesting. In other words, individual growth trajectories reside at L1, differences in growth rates across individuals within teams compose the L2 model, and the variation across teams is the L3 model (Raudenbush & Bryk, 2002, chap. 8).

Assuming that the three levels are individuals, teams, and organizations, a three-level model requires a subscript *k* to distinguish various organizations. For instance, X_{ijk} is the LMX score for the *i*th individual who is a member of team *j* within organization *k*. Similarly, W_{jk} is the leadership climate score for team *j* in organization *k*, and Q_k is the organizational culture value for organization *k*. This inclusion of a third-level variable also entails the addition of an additional residual value: (ν_{0k}). This implies the potential of additional third-level variance components: intercept variance (φ_{00}), slope variance (φ_{11}), and intercept-slope covariance (φ_{01}).

This inclusion of additional variance components allows for several kinds of ICCs to be calculated (Hox, 2010; Snijders & Bosker, 2012). First, is the proportion of total variance explained by the L3 variable, which is $\varphi_{00} / [\varphi_{00} + \tau_{00} + \sigma^2]$. Similar ICCs can be calculated for each level. Second, is the proportion of total variance explained by the L3 and L2 variables, which is $[\varphi_{00} + \tau_{00}] / [\varphi_{00} + \tau_{00} + \sigma^2]$. Third, is the proportion of variance shared by the L3 and L2 variables, which is $\varphi_{00} / [\varphi_{00} + \tau_{00}]$.

In addition to a variety of ICCs, there are a number of different regressions that can be performed when conducting a three-level analysis. For example, for each L1 variable, there are three different types of regressions: within-L2 regression, within-L3/between-L2 regression, and between-L3 regression (Snijders & Bosker, 2012).

Testing for complex models, such as three-level cross-level interactions involves expanding Equation 16 to include all first-order effects, all two-way cross-level interaction effects, and finally the term carrying information about the three-level interaction effect: $\left(X_{ijk} - \overline{X}_{jk}\right)\left(W_{jk} - \overline{W}_k\right)\left(Q_k - \overline{Q}\right)$. All issues we discussed earlier regarding two-way cross-level interactions apply to the three-level interaction context (Hox, 2010; Snijders & Bosker, 2012). For example, model building and centering the variables can all be generalized from a two-level model. Furthermore, there should be a clear definition of the cross-level interaction effect, all constituent terms should be included in the equation, and so forth.

A challenge regarding tests of three-way cross-level interaction effects is that it will be necessary to collect data from multiple higher-level units to capture possible variability of

L1 and L2 intercepts and slopes across L3 units. In fact, in many cases researchers may abandon hypotheses involving three-level interaction effects due to insufficient evidence regarding variation in intercepts and/or slopes at the third level. For example, Raudenbush, Rowan, and Cheong (1993) conducted a study involving the following three levels: (L1) classes, (L2) teachers, and (L3) schools. However, "because the number of schools was small and because there was little evidence of school-to-school variation, no level-3 predictors were specified" (Raudenbush & Bryk, 2002: 237).

We are not aware of a tool that would allow for the estimation of statistical power to detect three-level interaction effects. Although Konstantopolous (2008a, 2008b) addressed statistical power in the context of three-level models, this work refers to statistical power computations specifically for a dummy-coded treatment effect (i.e., main effect), but it does not address computations regarding three-level cross-level interaction effects. Nevertheless, given the increased level of complexity of the model tested and results regarding the importance of the lower-level sample size regarding power reported by Mathieu et al. (2012), other things equal, statistical power for detecting a three-level interaction effect is unlikely to be greater than the power to detect a two-level cross-level interaction effect. Thus, our recommendation is to use the Mathieu et al. power calculator in making research design decisions to make sure there is sufficient power to detect each of the two-level cross-level interaction effects. Although this will not guarantee sufficient power, this will at least produce some evidence about the probability of detecting a three-level interaction effect. Moreover, there should be a strong theory-based rationale to posit such a complex interaction effect. Clearly, there is a need for future work regarding the statistical power of the three-level interaction effect test.

Issue 9: What is the practical significance of the cross-level interaction effect? An issue also related to the interpretation of results refers to the practical significance of a cross-level interaction effect. A necessary step for understanding the practical significance of the cross-level interaction effect is to estimate the strength of the effect (Aguinis, Werner, Abbott, Angert, Park, & Kohlhausen, 2010). When using OLS regression, researchers usually estimate effect sizes based on the extent to which a variable predicts outcomes of interest (i.e., regression coefficient associated with the product term) or based on fit (i.e., proportion of variance explained by the interaction effect, usually assessed using R^2; Aguinis, 2004). Similar options are available in the context of multilevel modeling, and each one has advantages and disadvantages. Next, we describe each of these options by relying mainly on work by Hox (2010), Roberts, Monaco, Stovall, and Foster (2011), and Snijders and Bosker (2012).

The first option is to focus on the extent to which the cross-level interaction predicts the outcome of interest, which is indicated by γ_{11}. This is a useful indicator because it refers to the original metric used in collecting the data. However, the other side of the coin is that, precisely because the coefficient is scale specific, its size depends on the measures used to assess *X, Y*, and *W*. For example, referring back to our illustration, if a researcher uses a 100-point scale for empowerment, the resulting cross-level interaction effect γ_{11} will be much larger than if a researcher uses a 7-point scale. Because γ_{11} provides information regarding the prediction of *Y* scores, it is considered an index of an interaction's *predictive power*.

A second option for assessing effect size that has the advantage of scale-independence consists of focusing on the cross-level interaction's *explanatory power*: the proportion of the total variability of the slope of *Y* on *X* across teams that is explained by the L2 predictor *W*. To do so, we refer back to Equation 16, in which u_{1j} is the error term and its variance, denoted by τ_{11}, represents the total across-team variance in slopes. Equation 15 shows also the error term u_{1j} (i.e., the portion of β_{1j} that is independent of W_j). Note that u_{1j} in Equation 16 is what is left unexplained after controlling for the effect of *W*, and we use the symbol τ_{11w} to refer to the variance of this error term. Accordingly, we can calculate the proportion of total across-team variance in slopes explained specifically by *W* as follows:

$$\frac{\tau_{11} - \tau_{11w}}{\tau_{11}} \tag{22}$$

We computed the proportion of the total slope variance explained by the moderating effect of leadership climate using results shown in Table 1. We found that = $\frac{\tau_{11} - \tau_{11w}}{\tau_{11}}$ = $\frac{.025 - .019}{.025}$ = .24. In other words, *W* accounts for 24% of the total variance of β_{1j} across teams. This is a useful indicator of practical significance because it can be used to understand the relative importance of effects within one study and also across studies given that the metric is proportion of variance explained.

A third, commonly used, option is to estimate the effect size using a "pseudo R^2" metric. In multilevel modeling, we can obtain a pseudo R^2 value for each of the steps in the model building process, which we have done and reported in Table 1 using our illustrative data. For example, for Step 2, which involves the RIFSM, predicted criterion scores are obtained as follows,

$$\widehat{Y}_{ij} = \gamma_{00} + \gamma_{10}\left(X_{ij} - \overline{X}_j\right) + \gamma_{01}\left(W_j - \overline{W}\right) \tag{23}$$

which is the same as Equation 8, but excluding the error terms u_{0j} and r_{ij}. As shown in Table 1, pseudo R^2 increased from no variance explained by the null model to 23% of variance explained by the RIFSM. In other words, the addition of the coefficient associated with the L2 predictor increased variance explained by another 23%. The computation of pseudo R^2 for Step 3, which involves the RIRSM, involves calculating the squared correlation between observed and predicted Y_{ij} scores based on Equation 12 and excluding the error terms $u_{0j}, u_{1j}\ (X_{ij} - \overline{X}_j)$, and r_{ij} as follows:

$$\widehat{Y}_{ij} = \gamma_{00} + \gamma_{01}\left(W_j - \overline{W}\right) + \gamma_{10}\left(X_{ij} - \overline{X}_j\right) \tag{24}$$

Note that, because we do not use the variance component terms in predicting Y_{ij} scores, Equations 23 and 24 are identical, although they predict Y_{ij} scores for different steps in the model building process. This is why pseudo R^2 values are nearly identical for Steps 2 and 3, although a comparison of Equations 8 and 12 shows that these models are quite different.

The exclusion of variance components from the computation of pseudo R^2 values explains why some results can be counterintuitive, such as pseudo R^2 values becoming smaller when predictors are added to the model. Thus, this is the reason why Snijders and Bosker (2012: 109) noted that the computation of pseudo R^2 values "now and then leads to unpleasant surprises."

Table 1 also shows that the addition of the cross-product term in Step 4 leads to an increase of about 1% of variance explained. Once again, however, note that predicted Y_{ij} scores are obtained using an equation that excludes terms involving variance components as follows (which is Equation 16 without the variance component terms, $u_{0j}, u_{1j}(X_{ij} - \overline{X}_j)$, and + r_{ij}):

$$\widehat{Y}_{ij} = \gamma_{00} + \gamma_{01}\left(W_j - \overline{W}\right) + \gamma_{10}\left(X_{ij} - \overline{X}_j\right) + \gamma_{11}\left(X_{ij} - \overline{X}_j\right)\left(W_j - \overline{W}\right) \tag{25}$$

So, this result means that there is an additional 1% of variance explained by adding the γ_{11} coefficient to the model, but there is no information regarding variance components and their effects on the proportion of variance explained in Y_{ij} scores. A primary advantage of multilevel modeling is the decomposition of various sources of variance based on the level at which each source of variance resides. However, the computation of pseudo R^2 values does not take these different sources of variance into account. In other words, pseudo R^2 values are based on the fixed portion of the models only and ignore the random terms. This is why "the estimated values for R^2 usually change only very little when random regression coefficients are included in the model" (Snijders & Bosker, 2012: 113). Another weakness to this approach is that there is the potential for one to obtain a negative pseudo R^2 value, but this likely means that the model is misspecified (Hox, 2010; Snijders & Bosker, 2012). In sum, although we report pseudo R^2 values in Table 1 and our annotated R code includes the appropriate commands for all computations, it is important to understand the meaning and interpretation of these values specifically in the context of multilevel modeling.

In sum, each of the three options we described for reporting effect sizes and interpreting the practical significance of a cross-level interaction effect has advantages and disadvantages. So, our recommendation is that researchers report all three, together with statements about how each one should be interpreted. This recommendation follows the principle of full disclosure and, following a customer-centric approach (Aguinis et al., 2010), also allows readers the opportunity to interpret the meaningfulness of results themselves. Moreover, also related to the customer-centric approach to reporting significant results (Aguinis et al., 2010), we emphasize that effect sizes should be interpreted within specific contexts and the fact that if an effect seems small in terms of the proportion of variance explained, it does not automatically mean that it is unimportant in terms of theory or practice.

Issue 10: What information should be reported based on multilevel modeling analyses? The field of management lacks clear reporting standards regarding multilevel modeling. There is wide variability in terms of the type of information that researchers choose to present in their tables—and how that information is presented. In contrast, the American Psychological Association is quite clear regarding what type of information should be reported when a study includes popular and long-established techniques such as multiple regression and ANOVA (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008). Although the 250-plus-page APA *Publication Manual* does not refer to multilevel modeling at all, it does include a "Sample Multilevel Model Table" (Table 5.15) that can be used when reporting multilevel modeling results (American Psychological Association, 2009: 147-148). Unfortunately, the APA *Publication Manual* does not include any text or rationale for why each piece of information should be included in this table, and, moreover, the proposed template is not sufficiently comprehensive. For example, the APA template does not include information on ICC, number of estimated parameters, and pseudo R^2. As mentioned earlier, ICC information is needed for readers to understand whether the use of multilevel modeling was in fact justified. Including the number of estimated parameters is also useful so that readers can quickly and accurately understand the nature of the model. Also as mentioned earlier, pseudo R^2 information is also a useful, albeit imperfect, effect size metric. In addition, our proposed table is also more comprehensive than its APA counterpart because it includes sample size and clear labels for each model. Overall, reporting the information included in Table 1 is important because, absent this information, results based on multilevel modeling can be perceived as lacking transparency.

Our Table 1 can be used as a template for the type of information that needs to be reported when conducting a multilevel study regardless of the particular focus—L1 direct effects, L2 direct effects, cross-level interactions. This table includes clear labels regarding which are the variables at which level and the sample size for each level as well as the coefficients for each effect—including their standard errors and statistical significance. This table also includes a crucial piece of information that is often missing from published multilevel research: complete information regarding the size of each variance component. This information is important for several reasons. First, as statistical software programs become increasingly available and easy to use, there are instances in which users may not fully understand the resulting output. Routinely reporting variance components will allow researchers to become more familiar with their data, results, and interpretation of results. Second, given increasing concerns about ethical violations and data "massaging" (Bedeian, Taylor, & Miller, 2010), reporting variance components can allow a skeptical scientific audience to double check results and possible instances of misreporting (either by error or intention). Overall, multilevel research can benefit from a greater degree of standardization and openness in terms of the information that is reported so our recommendations will be useful in this regard. Finally, Table 1 also shows all of this information for each of the steps in the model building process.

As multilevel modeling becomes a more popular approach in management and related fields, it is important that results from such analyses be reported in a detailed and comprehensive manner. Such clear and standardized reporting serves several purposes. First, it allows readers to have all the necessary information to fully understand and interpret results. Second, it allows for the possibility that future research can replicate the results of any one particular study. Third, it allows for the possibility of making results more useful and accurate in terms of their future inclusion in subsequent literature reviews, both qualitative and quantitative (e.g., meta-analysis). Finally, the availability of information regarding the variability of slopes across groups allows for a more precise computation of statistical power, which is particularly important in cases when evidence seems to suggest the absence of a cross-level interaction effect.

Concluding Comments

Understanding the interplay between variables at different levels of analysis is a key substantive challenge in management research (Bliese, 2000). Thus, there is an increased interest in multilevel models—models that include variables at more than one level of analysis. For example, a review by Aguinis, Pierce, Bosco, and Muslin (2009) revealed that multilevel modeling was the third most popular data-analytic approach in articles published in *Organizational Research Methods* from 1998 to 2007 (behind multiple regression/correlation and structural equation modeling). Moreover, Aguinis et al. (2009) documented that multilevel modeling has gained more popularity than any other data-analytic approach over this time period. One reason for the increased popularity is that such models allow for the assessment of whether relationships among entities are moderated by variables at the collective level within which these entities reside. Given the nature of organizational life in the 21st century, and the fact that people work in increasingly interdependent environments, shared influences including leadership, policies, practices, and many other processes create dependence in the data regardless of whether nesting is formally established through organizational structures (Cascio & Aguinis, 2008). Thus, data nonindependence is likely to be more pervasive than typically acknowledged.

Our review of questions posted on listservs suggests that researchers are mostly concerned with issues related to data analysis and interpretation of results. In other words, researchers are more concerned with and interested in answering questions about how to handle the data that have been collected compared to how to plan and execute future data collection efforts. A similar emphasis on data analysis issues relative to research design and measurement issues has been documented by other reviews. For example, Aguinis et al. (2009: 106) concluded that “an implication of our study is that more attention is needed regarding the development of new as well as the improvement of existing research designs.” Although our article provides recommendations regarding actions researchers can take before and after data are collected, we believe that the most impactful decisions take place during the early stages of research including conceptualization, research design, and measurement. If models are misspecified (e.g., important L2 variables are not included in the study), research design is suboptimal (e.g., sample size is too small to detect existing cross-level interactions), and measures are not reliable (i.e., leading to measurement error), then issues around interpretation become less relevant because they can turn into attempts to fix unfixable design and measurement problems. In closing, drawing meaningful and accurate conclusions about cross-level interaction effects involves important decision points, and we hope our article will be a useful resource in this regard.

Appendix A

Annotated R Code for Multilevel Analysis With Illustrative Data

Note: also available at http://mypage.iu.edu/~haguinis.

```
#Setting Working Directory and Reading Data File
library('lme4')
library('RLRsim')
setwd('C:/Documents/JOM')
exdata=read.csv('JOM.csv')

#STEP 1: Null Model
lmm.fit1=lmer(Y ~(1|l2id),data=exdata,REML=F)
summary(lmm.fit1)

  # Compute ICC
  iccy=VarCorr(lmm.fit1)$l2id[1,1]/(VarCorr(lmm.fit1)$l2id
  [1,1]+attr(VarCorr(lmm.fit1),'sc')^2)]
  iccy

#STEP 2: Random Intercept and Fixed Slope Model
lmm.fit2=lmer(Y ~(1|l2id)+Xc+I(Wj-mean(Wj) ),
data=exdata,REML=F)
summary(lmm.fit2)

  # Computing pseudo R-squared
  yhat2=model.matrix(lmm.fit2)%*%fixef(lmm.fit2)
  cor(yhat2,exdata$Y)^2

#STEP 3: Random Intercept and Random Slope model
lmm.fit3=lmer(Y ~Xc+(Xc|l2id)+I(Wj-mean(Wj) ),
data=exdata,REML=F)
summary(lmm.fit3)

  # Print VC Estimates
  VarCorr(lmm.fit3)

  # Computing pseudo R-squared
  yhat3=model.matrix(lmm.fit3)%*%fixef(lmm.fit3)
  cor(yhat3,exdata$Y)^2
  # Crainceanu & Ruppert (2004) Test of Slope Variance
  Component
  obs.LRT <- 2*(logLik(lmm.fit3)-logLik(lmm.fit2))[1]
  X <- lmm.fit3@X
  Z <- t(as.matrix(lmm.fit3@Zt))
  sim.LRT <- LRTSim(X, Z, 0, diag(ncol(Z)))
  (pval <- mean(sim.LRT > obs.LRT))

  # Nonparametric Bootstrap Function
  REMLVC=VarCorr(lmer(Y ~Xc+(Xc|l2id)+I(Wj-mean(Wj)
  ),data=exdata,REML=T))$l2id[1:2,1:2] U.R=chol(REMLVC)
  REbootstrap=function(Us,es,X,gs){
  nj=nrow(Us)
  idk=sample(1:nj,size=nj,replace=T)
  Usk=as.matrix(Us[idk,])
  esk=sample(es,size=length(es),replace=T)
  S=t(Usk)%*%Usk/nj
  U.S = chol(S)
  A=solve(U.S)%*%U.R
  Usk = Usk%*%A
  datk=expand.grid(l1id = 1:6,l2id = 1:nj)
  colnames(X)=c(‘one’,‘Xc’,‘Wjc’)
  datk=cbind(datk,X)
  datk$yk = X%*%gs + Usk[datk$l2id,1]+Usk[datk$l2id,2]*X[,2
  ]+esk
  lmm.fitk=lmer(yk ~Xc+(Xc|l2id)+Wjc,data=datk,REML=F)
  tau11k = VarCorr(lmm.fitk)$l2id[2,2]
  tau11k
  }

  # Implementing Bootstrap
  bootks=replicate(1500,REbootstrap(Us=ranef(lmm.
  fit3)$l2id,es=resid(lmm.fit3),X=model.matrix(lmm.
  fit3),gs=fixef(lmm.fit3)))quantile(bootks,prob
  s=c(.025,.975))

#STEP 4: Cross-Level Interaction Model
lmm.fit4=lmer(Y ~(Xc|l2id)+Xc*I(Wj-mean(Wj) ),
data=exdata,REML=F)
summary(lmm.fit4)

  # Print VC Estimates
  VarCorr(lmm.fit4)

  # Computing pseudo R-squared
  yhat4=model.matrix(lmm.fit4)%*%fixef(lmm.fit4)
  cor(yhat4,exdata$Y)^2
```

```
#Interaction Plots
#Code creates graphs in pdf format in the same directory
as the data file
gammas=fixef(lmm.fit4)

pdf(‘intplot.xw.pdf’,width=10,height=8)
par(mar=c(3.25,3.25,.5,.5),cex=2,bty=‘l’,las=1,family=‘seri
  f’,mgp=c(1.85,.5,0))
#Figure 3 Panel (a) - Full Y Scale
Wjs=c(0-sd(exdata$Wj),0,0+sd(exdata$Wj))
xlb=mean(exdata$Xc)-sd(exdata$Xc);xub=mean(exdata$Xc)+sd(ex
data$Xc)
ylb=1;yub=7
curve(0+1*x,xlb,xub,xlab=‘LMX’,ylab=‘Individual
  Empowerment’,lwd=2,type=‘n’,
ylim=c(ylb,yub))
for(i in 1:length(Wjs)){
B0j=gammas[1]+gammas[3]*Wjs[i]
B1j=gammas[2]+gammas[4]*Wjs[i]
curve(B0j+B1j*x,xlb,xub,add=T,xlab=‘LMX’,ylab=‘Individual
  Empowerment’,lwd=2,lty=i)
}
labs=c(expression(W[j]==-1*~~SD),expression(W[j]==0*~~SD),
  expression(W[j]==1*~~SD))
legend(xlb,5,legend=c("Leadership Climate",labs[1],labs[2],
  labs[3]),bty=‘n’,lty=c(0:3))

#Figure 3 Panel (b) - Reduced Y Scale
ylb=5;yub=6.5
curve(0+1*x,xlb,xub,xlab=‘LMX’,ylab=‘Individual
  Empowerment’,lwd=2,type=‘n’,
ylim=c(ylb,yub))
for(i in 1:length(Wjs)){
B0j=gammas[1]+gammas[3]*Wjs[i]
B1j=gammas[2]+gammas[4]*Wjs[i]
curve(B0j+B1j*x,xlb,xub,add=T,xlab=‘LMX’,ylab=‘Individual
  Empowerment’,lwd=2,lty=i)
}
labs=c(expression(W[j]==-1*~~SD),expression(W[j]==0*~~SD),
  expression(W[j]==1*~~SD))
legend(xlb,6.5,legend=c("Leadership Climate",labs[1],
  labs[2],labs[3]),bty=‘n’,lty=c(0:3))
dev.off()
```

Appendix B

Annotated R Code for Power Analysis From Mathieu et al. (2012) Using This Article's Illustrative Data

Note: also available at http://mypage.iu.edu/~haguinis.

```
l2n = 105     #Level-2 sample size
l1n = 6       #Average Level-1 sample size
iccx = .12      #ICC1 for X
g00 = 0       #Intercept for B0j equation (Level-1 intercept)
g01 = 0       #Direct cross-level effect of average Xj on Y
g02 = 0       #Direct cross-level effect of W on Y
g03 = 0       #Between-group  interaction  effect  between
              W and Xj on Y
g10 = 0.4     #Intercept  for  B1j  equation  (Level-1  effect
              of X on Y)
g11 = 0.15      #Cross-level interaction effect
vu0j = 0.01     #Variance component for intercept
vu1j = 0.1      #SD of Level-1 slopes
vresid = 0.8  #Variance component for residual, within variance
alpha = .05     #Rejection level
REPS = 1000     #Number  of  Monte  Carlo  Replications,  1000
                recommended

hlmmmr <-
function(iccx,l2n,l1n,g00,g01,g02,g03,g10,g11,vu0j,vu1j,al
pha){
require(lme4.0)
Wj = rnorm(l2n, 0, sd=1)
Xbarj = rnorm(l2n, 0, sd=sqrt(iccx)) ## Level-2 effects on x
b0=g00+g01*Xbarj+g02*Wj+g03*Xbarj*Wj+rnorm(l2n,0,sd=sqrt(vu0j))
b1 = g10 + g11*Wj + rnorm(l2n,0,sd=sqrt(vu1j))
dat=expand.grid(l1id = 1:l1n,l2id = 1:l2n)
dat$X=rnorm(l1n*l2n,0,sd=sqrt(1-iccx))+Xbarj[dat[,2]]
dat$Xbarj=Xbarj[dat[,2]]
dat$Wj=Wj[dat[,2]]
dat$Y <- b0[dat$l2id]+ b1[dat$l2id]*(dat$X-dat$Xbarj)+rnorm
(l1n*l2n,0,sd=sqrt(vresid))
dat$Xc=(dat$X - Xbarj[dat[,2]])
lmm.fit<- lmer(Y ~ Xc+Xbarj+Wj+Xbarj:Wj+Xc:Wj+(Xc|l2id),data=
dat)
fe.g <- fixef(lmm.fit)
fe.se <- sqrt(diag(vcov(lmm.fit)))
ifelse(abs(fe.g[6]/fe.se[6])>qt(1-alpha/2,l2n-4),1,0)
}
simout=replicate(REPS,hlmmmr(iccx,l2n,l1n,g00,g01,g02,g03,g10,
g11,vu0j,vu1j,alpha))
powerEST=mean(simout)
powerEST
```

Note

[1] As noted by an anonymous reviewer, ICC greater than 0 implies that within-team dependence must be taken into account in computing standard errors. But ICC greater than 0 does not necessarily mean that there is a need to model the effect of *W* on β_{0j} unless there is an interest in *W* as a substantive or control variable. However, as we describe later in our article, it is necessary to model *W* when there is an interest in a cross-level interaction between *X* and *W*.

References

Aguinis, H. 2004. *Regression analysis for categorical moderators*. New York: Guilford.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. 2005. Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology*, 90: 94-107.

Aguinis, H., Boyd, B. K., Pierce, C. A., & Short, J. C. 2011. Walking new avenues in management research methods and theories: Bridging micro and macro domains. *Journal of Management*, 37: 395-403.

Aguinis, H., Gottfredson, R. K., & Wright, T. A. 2011. Best-practice recommendations for estimating interaction effects using meta-analysis. *Journal of Organizational Behavior*, 32: 1033-1043.

Aguinis, H., Pierce, C. A., Bosco, F. A., & Muslin, I. S. 2009. First decade of *Organizational Research Methods*: Trends in design, measurement, and data-analysis topics. *Organizational Research Methods*, 12: 69-112.

Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H., & Kohlhausen, D. 2010. Customer-centric science: Reporting research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods*, 13: 515-539.

Aiken, L. S., & West, S. G. 1991. *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

American Psychological Association. 2009. *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. 2008. Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, 63: 839-851.

Aytug, Z. G., Rothstein, H. R., Zhou, W., & Kern, M. C. 2012. Revealed or concealed? Transparency of procedures, decisions, and judgment calls in meta-analyses. *Organizational Research Methods*, 15: 103-133.

Barney, J. B., Ketchen, D. J., & Wright, M. 2011. The future of resource-based theory: Revitalization or decline? *Journal of Management*, 37: 1299-1315.

Bedeian, A. G., Taylor, S. G., & Miller, A. N. 2010. Management science on the credibility bubble: Cardinal sins and various misdemeanors. *Academy of Management Learning and Education*, 9: 715-725.

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

Bliese, P. D. 2002. Multilevel random coefficient modeling in organizational research: Examples using SAS and S-PLUS. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis*: 401-445. San Francisco: Jossey-Bass.

Bliese, P. D., & Hanges, P. J. 2004. Being both too liberal and too conservative: The perils of treating grouped data as though they were independent. *Organizational Research Methods*, 7: 400-417.

Bosker, R. J., Snijders, T. A. B., & Guldemond, H. 2003. *PINT (Power IN Two-level designs): Estimating standard errors of regression coefficients in hierarchical linear models for power calculations* (Version 2.1). Groningen, Netherlands: Rijksuniversiteit Groningen.

Brutus, S., Aguinis, H., & Wassmer, U. 2013. Self-reported limitations and future directions in scholarly reports: Analysis and recommendations. *Journal of Management*, 39: 48-75.

Carpenter, J. R., Goldstein, H., & Rasbash, J. 2003. A novel bootstrap procedure for assessing the relationship between class size and achievement. *Applied Statistics*, 52: 431-443.

Cascio, W. F., & Aguinis, H. 2008. Staffing twenty-first-century organizations. *Academy of Management Annals*, 2: 133-165.

Chen, G., Kirkman, B. L., Kanfer, R., Allen, D., & Rosen, B. 2007. A multilevel study of leadership, empowerment, and performance in teams. *Journal of Applied Psychology*, 92: 331-346.

Covin, J. G., & Wales, W. J. 2012. The measurement of entrepreneurial orientation. *Entrepreneurship Theory and Practice*, 36: 677-702.

Crainiceanu, C., & Ruppert, D. 2004. Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society-B*, 66: 165-185.

Culpepper, S. A. 2010. Studying individual differences in predictability with gamma regression and nonlinear multilevel models. *Multivariate Behavioral Research*, 45: 153-185.

Culpepper, S. A., & Aguinis, H. 2011. Using analysis of covariance (ANCOVA) with fallible covariates. *Psychological Methods*, 16: 166-178.

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: 339-362.

Dalton, D. R., & Aguinis, H. 2013. Measurement malaise in strategic management studies: The case of corporate governance research. *Organizational Research Methods*, 16: 88-99.

Dalton, D. R., Aguinis, H., Dalton, C. A., Bosco, F. A., & Pierce, C. A. 2012. Revisiting the file drawer problem in meta-analysis: An empirical assessment of published and non-published correlation matrices. *Personnel Psychology*, 65: 221-249.

Davison, M. L., Kwak, N., Seo, Y. S., & Choi, J. 2002. Using hierarchical linear models to examine moderator effects: Person-by-organization interactions. *Organizational Research Methods*, 5: 231-254.

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

Field, C. A., & Welsh, A. H. 2007. Bootstrapping clustered data. *Journal of the Royal Statistical Society-B*, 69: 369-390.

Gelfand, A. E., Sahu, S. K., & Carlin, B. P. 1995. Efficient parametrisations for normal linear mixed models. *Biometrika*, 82: 479-488.

Han, J. 2005. Crossover linear modeling: Combining multilevel heterogeneities in crossover relationships. *Organizational Research Methods*, 8: 290-316.

Hedges, L. V., & Hedberg, E. C. 2007. Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29: 60-87.

Hofmann, D. A., & Gavin, M. B. 1998. Centering decisions in hierarchical linear models: Theoretical and methodological implications for organizational science. *Journal of Management*, 24: 623-641.

Hox, J. 2010. *Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge.

Kahn, J. H. 2011. Multilevel modeling: Overview and applications to research in counseling psychology. *Journal of Counseling Psychology*, 58: 257-271.

Kenny, D. A., & Judd, C. M. 1996. A general procedure for the estimation of interdependence. *Psychological Bulletin*, 119: 138-148.

Konstantopolous, S. 2008a. The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness*, 1: 265-288.

Konstantopolous, S. 2008b. The power of the test for treatment effects in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness*, 1: 66-88.

Kozlowski, S. W. J., & Klein, K. J. 2000. A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions*: 3-90. San Francisco: Jossey-Bass.

Kreft, I., & De Leeuw, J. 1998. *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

Liden, R. C., & Antonakis, J. 2009. Considering context in psychological leadership research. *Human Relations*, 62: 1587-1605.

Maas, C. J. M., & Hox, J. J. 2004. The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*, 46: 427-440.

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. 2012. Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology*, 97: 951-966.

Mathieu, J. E., & Chen, G. 2011. The etiology of the multilevel paradigm in management research. *Journal of Management*, 37: 610-641.

Maxwell, S. E. 2004. The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, 9: 147-163.

Molloy, J. C., Ployhart, R. E., & Wright, P. M. 2011. The myth of "the" micro-macro divide: Bridging system-level and disciplinary divides. *Journal of Management*, 37: 581-609.

Peugh, J. L. 2010. A practical guide to multilevel modeling. *Journal of School Psychology*, 48: 85-112.

Pierce, J. R., & Aguinis, H. 2013. The too-much-of-a-good-thing effect in management. *Journal of Management*, 39: 313-338.

Preacher, K. J., Curran, P. J., & Bauer, D. J. 2006. Computational tools for probing interaction effects in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, 31: 437-448.

Preacher, K. J., Zyphur, M. J., & Zhang, Z. 2010. A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, 15: 209-233.

Raudenbush, S. W. 1997. Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2: 173-185.

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

Raudenbush, S. W., & Liu, X. 2000. Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, 5: 199-213.

Raudenbush, S. W., Rowan, B., & Cheong, Y. F. 1993. The pursuit of higher-order instructional goals in secondary schools: Class, teacher, and school influences. *American Educational Research Journal*, 30: 523-553.

Roberts, J. K., Monaco, J. P., Stovall, H., & Foster, V. 2011. Explained variance in multilevel models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis*: 219-230. New York: Routledge.

Scheipl, F., Greven, S., & Kuechenhoff, H. 2008. Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis*, 52: 3283-3299.

Scherbaum, C. A., & Ferreter, J. M. 2009. Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods*, 12: 347-367.

Singer, J. D. 1998. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 24: 323-355.

Snijders, T. A. B. 2005. Power and sample size in multilevel linear models. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science*, vol. 3: 1570-1573. Chichester, UK: Wiley.

Snijders, T. A. B., & Bosker, R. J. 2012. *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Los Angeles, CA: Sage.

Spybrook, J., Bloom, H., Congdon, R., Hill, C. Martinez, A., & Raudenbush, S. W. 2011. Optimal design plus empirical evidence: Documentation for the "Optimal Design" software (version 3.0). Retrieved from http://www.wtgrantfoundation.org/resources/consultation-service-and-optimal-design.

Stram, D., & Lee, J. W. 1994. Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50: 1171-1177.

Uy, M. A., Foo, M. D., & Aguinis, H. 2010. Using event sampling methodology to advance entrepreneurship theory and research. *Organizational Research Methods*, 13: 31-54.

Wolfinger, R., & O'Connell, M. 2007. Generalized linear mixed models: A pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48: 233-243.