Journal of Applied Psychology
2012, Vol. 97, No. 5, 951–966

0021-9010/12/$12.00 DOI: 10.1037/a0028380

Understanding and Estimating the Power to Detect Cross-Level Interaction Effects in Multilevel Modeling

John E. Mathieu
University of Connecticut

Herman Aguinis
Indiana University

Steven A. Culpepper
University of Illinois at Urbana–Champaign

Gilad Chen
University of Maryland

Cross-level interaction effects lie at the heart of multilevel contingency and interactionism theories. Researchers have often lamented the difficulty of finding hypothesized cross-level interactions, and to date there has been no means by which the statistical power of such tests can be evaluated. We develop such a method and report results of a large-scale simulation study, verify its accuracy, and provide evidence regarding the relative importance of factors that affect the power to detect cross-level interactions. Our results indicate that the statistical power to detect cross-level interactions is determined primarily by the magnitude of the cross-level interaction, the standard deviation of lower level slopes, and the lower and upper level sample sizes. We provide a Monte Carlo tool that enables researchers to a priori design more efficient multilevel studies and provides a means by which they can better interpret potential explanations for nonsignificant results. We conclude with recommendations for how scholars might design future multilevel studies that will lead to more accurate inferences regarding the presence of cross-level interactions.

Keywords: multilevel modeling, interactions, power, interactionism

The past quarter century or so has witnessed a growing application of multilevel theories, designs, and analyses in applied psychology (cf. Aguinis, Boyd, Pierce, & Short, 2011; Bliese, Chan, & Ployhart, 2007; Griffin, 2007; Hitt, Beamish, Jackson, & Mathieu, 2007; Klein, Cannella, & Tosi, 1999; Mathieu & Chen, 2011; Rousseau, 1985). The essence of the multilevel approach is that an outcome of interest is conceptualized as resulting from a combination of influences emanating from the same level as well as higher levels of analysis. Moreover, the multilevel approach formally recognizes that entities (e.g., individuals) are typically nested in higher level collectives (e.g., teams, organizations), which leads to nontrivial theoretical and analytical implications. We conducted a systematic review of all articles published in the *Journal of Applied Psychology* from 2000 to 2010, and results show that multilevel studies have increased in frequency rapidly. In the years 2000–2002, there was an average of three multilevel articles published in the *Journal of Applied Psychology* per year, which has risen steadily to an average of 13 articles per year over the 2008–2010 period. Aguinis, Pierce, Bosco, and Muslin's (2009) recent review of articles published in *Organizational Research Methods* between 1998 and 2007 has also documented a steady increase in the attention paid to multilevel methodology in organizational research.

The multilevel approach advances three types of relationships.[1] First, there are potential lower level direct influences, such as between individuals' knowledge, skills, abilities, and other characteristics (KSAOs) and their individual job performance. Second, there may be direct cross-level influences, such as the effects of group cohesion on group members' average job performance. And third, there may well be cross-level interactions whereby the relationships between lower level predictors and outcomes differ as a function of higher level factors. For example, the relationship between individuals' need for affiliation and their job performance might be accentuated to the extent that they are members of more cohesive groups. In sum, the multilevel approach in applied psychology has energized the examination of joint influences of predictors from different levels on lower level outcomes of interest while simultaneously recognizing the fact that in organizational settings, individuals are typically nested in higher level units and, hence, are amenable to contextual influences.

Although multilevel investigations are drawing increased attention to cross-level interactions, these are not really new phenomena. In fact, arguably what we now refer to as cross-level interactions were the focus of a debate concerning person versus situational influences

This article was published Online First May 14, 2012.

John E. Mathieu, Department of Management, University of Connecticut; Herman Aguinis, Kelley School of Business, Indiana University; Steven A. Culpepper, Department of Statistics, University of Illinois at Urbana–Champaign; Gilad Chen, Robert H. Smith School of Business, University of Maryland.

An earlier version of this work was presented in J. Cortina (Chair), "Independence Day? New Developments in Dealing with Nested Data," a symposium at the meeting of the Society of Industrial and Organizational Psychology, Atlanta, Georgia, April 2010. We contributed equally to this work, and the order of authorship was determined randomly.

Correspondence concerning this article should be addressed to John E. Mathieu, Department of Management, School of Business, University of Connecticut, 2100 Hillside Road, Unit 1041MG, Storrs, CT 06269-1041. E-mail: John.Mathieu@business.uconn.edu

[1] We recognize that multilevel approaches may also consider upward influences. However, we restrict our consideration to downward cross-level effects in this article.

on individuals' behaviors (e.g., Bowers, 1973; Endler & Magnusson, 1976; see Mischel, 2004, and Pervin, 1989, for reviews). Although varying in specific form, the consensus arising from that exchange was that individuals' behaviors were the result of some form of interaction between individual and situational forces (Buss, 1977; Mischel, 2004; Pervin, 1989; Schneider, 1983).

Historically, interactionism has usually been approached from analysis of variance (ANOVA) and moderated multiple regression perspectives, where situations were seen as categorical groupings or conditions that interact with individual differences. What the multilevel approach has contributed to this understanding is that because individuals are nested in higher level units (e.g., groups), this nesting changes the way in which situational factors should be conceptualized and analyzed. Rather than viewing situations as categorical differences (or treatments), the multilevel analytical approach can examine the influences, both direct and interactive, of continuous higher level variables on lower level outcomes. Of note, multilevel modeling has also been referred to as hierarchical linear modeling (Raudenbush & Bryk, 2002), mixed-effect models (Cao & Ramsay, 2010), random coefficient modeling (Longford, 1993), and covariance components models (e.g., Searle, Casella, & McCulloch, 1992). In this analytical paradigm, cross-level interactions lie at the heart of modern-day contingency theories, person–environment fit models, and any theory that considers outcomes to be a result of combined influences emanating from different levels of analysis (e.g., Grizzle, Zablah, Brown, Mowen, & Lee, 2009; Wallace, Edwards, Arnold, Frazier, & Finch, 2009; Yu, 2009).

We conducted a review of published *Journal of Applied Psychology* studies involving tests of cross-level interactions, and results suggest that researchers have typically considered the influence of three types of unit-level moderators: (a) unit-level climate; (b) ambient leadership practices directed at the unit; and (c) other unit-level emergent states such as collective efficacy, team empowerment, or conflict. Representative studies from the 2000–2010 decade are summarized in Table 1 (note that this table includes additional information regarding each study, including power analysis computations that we describe later in this article). An example of the first type of cross-level moderator is a study by Hofmann, Morgeson, and Gerras (2003), who found that the individual-level relationship between the quality of leader–member exchange (LMX) and the extent to which soldiers incorporated safety into their role definitions were more positive to the extent that unit-level safety climate was more positive. Similarly, Liao and Rupp (2005) examined how various aspects of justice climates moderate relationships between individual-level justice orientations and job attitudes. Illustrative of the second type of moderator, Mathieu, Ahearne, and Taylor (2007) found that employees' technology self-efficacy related more positively to their actual use of technology when leaders engaged in more empowering leadership behaviors toward their units. An example of the third type of cross-level moderator was illustrated in a study by Bliese and Jex (1999), who found that work-related stressors (e.g., work overload) related less negatively to job attitudes (e.g., job satisfaction) when unit members shared higher levels of collective efficacy beliefs. Finally, a study by Chen, Kirkman, Kanfer, Allen, and Rosen (2007) incorporated the latter two types of cross-level moderators, in examining how team-level empowering leadership

Table 1
Selective Summary of Previous Journal of Applied Psychology *Studies Testing Cross-Level Interactions*

Reference	Variables	n_i	n_j	ρ_x	ρ_y	ρ_{xx}	ρ_{yy}	ρ_{ww}	γ_{10}	$\sqrt{\tau_{11}}$	γ_{1w}	$\gamma_{0\bar{x}}$	$\gamma_{0\bar{x}w}$	γ_{0w}	Power
Chen et al. (2007)	*X*: Leader–member exchange	7.18	62	.12	.00	.93	.88	.93	.45	.18	.15	.35	.18	.02	(α = .01) .42
	W: Empowering leadership														(α = .05) .64
	Y: Individual empowerment														(α = .10) .77
Chen et al. (2007)	*X*: Individual empowerment	7.18	62	.00	.28	.88	.97	.91	.16	.11	−.06	.35	−.39	.10	(α = .01) .09
	W: Team empowerment														(α = .05) .23
	Y: Individual performance														(α = .10) .37
Hofmann et al. (2003)	*X*: Leader–member exchange	3.76	25	.39	.30	.94	.98	.94	.39	.27	.48	.01	.15	.25	(α = .01) .59
	W: Safety climate														(α = .05) .80
	Y: Safety role definitions														(α = .10) .88
Liao & Rupp (2005)	*X*: Justice orientation	5.25	44	.11	.11	.85	.81	.83	.10	.14	.12	.17	−.02	.37	(α = .01) .14
	W: Organizational focus on PJ climate														(α = .05) .33
	Y: Satisfaction with organization.														(α = .10) .44
Liao & Rupp (2005)	*X*: Justice orientation	5.25	44	.11	.30	.85	.89	.83	.07	.24	.08	−.08	.30	.56	(α = .01) .06
	W: Supervisor focus on PJ climate														(α = .05) .19
	Y: Supervisor commitment														(α = .10) .29
Mathieu et al. (2007)	*X*: Work experience	2.67	221	.02	.00	.85	.81	.98	−.06	.00	−.06	−.23	−.14	−.23	(α = .01) .10
	W: Empowering leadership														(α = .05) .30
	Y: Technology self-efficacy														(α = .10) .39
Mathieu et al. (2007)	*X*: Technology use	2.67	221	.00	.04	.81	.74	.98	.07	.18	.06	.20	.12	.03	(α = .01) .14
	W: Empowering leadership														(α = .05) .32
	Y: Individual performance														(α = .10) .44
M		4.85	97	.11	.15	.87	.87	.91	.17	.15	.11	.11	.03	.16	(α = .01) .22
															(α = .05) .40
															(α = .10) .51

Note. n_i = average Level 1 *N*; n_j = Level 2 *N*; ρ_x = Level 1 *X* ICC; ρ_y = Level 1 *Y* ICC; ρ_{xx} = Level 1 *X* reliability; ρ_{yy} = Level 1 *Y* reliability; ρ_{ww} = Level 2 *W* reliability; γ_{10} = average Level 1 direct effect; $\sqrt{\tau_{11}}$ = *SD* of Level 1 slopes without predictors; γ_{1w} = cross-level interaction; $\gamma_{0\bar{x}}$ = direct effect of $\bar{X}_j$; $\gamma_{0\bar{x}w}$ = direct effect of $\bar{X}_j W_j$; γ_{0w} = direct effect of W_j; PJ = procedural justice; ICC = intraclass correlation coefficient.

moderated the relationship between LMX and individual-level (psychological) empowerment and how team-level empowerment moderated the relationship between psychological empowerment and individual (team member–level) performance.

The common feature in all these studies is that relationships between variables at the individual or other lower level differ as a function of contextual features represented by the upper level moderator. Of course, other cross-level moderators beyond the examples noted above have been examined in the literature, but the ones we described have been considered most frequently thus far by researchers in applied psychology and management. Given the increasing prevalence of cross-level interactions in applied psychology theorizing and research, we argue that it is now critical to develop a clearer understanding of what research design and measurement factors are most likely to enable researchers to reach sufficient statistical power levels to detect theoretically meaningful cross-level interactions.

The advent of multilevel designs and analyses has raised several unique challenges for researchers, not the least of which are concerns about statistical power in the context of nested arrangements. In other words, multilevel designs render traditional methods of estimating statistical power inapplicable, given the complex nonindependence of lower level observations (Snijders & Bosker, 1999). This has been particularly problematic for the study of cross-level interactions, for which researchers may have beliefs about the statistical power of their tests but no way of assessing the accuracy of such beliefs. For example, in a study of individual safety behavior, Neal and Griffin (2006) noted that "the power of this study to detect effects at the group level of analysis was limited. With only 33 work groups, we only had sufficient power to detect large effect sizes" (p. 952). Elsewhere, Grizzle et al. (2009) lamented, "We also note that our statistical power to identify significant cross-level and aggregate-level effects was limited. This limitation was due to both our relatively small unit-level sample size ($n = 38$) and moderately low group mean reliability for our unit climate measure, $ICC(2) = .61$" (p. 1238). Although such concerns are often voiced by multilevel scholars, to date there has not been a means by which they can calculate the actual power of their cross-level interaction tests (Scherbaum & Ferreter, 2009; Snijders & Bosker, 1999). Without being able to estimate the actual statistical power of cross-level interaction tests, researchers stand the risks of designing suboptimal multilevel studies, as well as erroneously concluding that meaningful substantive effects are nonexistent, both of which can potentially undermine important substantive discoveries.

Accordingly, in the present investigation we advance a method to calculate the power of cross-level interaction tests in multilevel studies. Below we first outline how power considerations apply to investigations of cross-level interactions. We then present a large-scale simulation study that will allow us to understand the relative impact of various research design and measurement factors on the power to detect cross-level interactions, and we also describe a tool that researchers can use to estimate such power a priori. We then estimate the power of previous empirical investigations to illustrate the utility of the new power estimator for substantive research in applied psychology. Finally, we conclude with recommendations for how scholars might design future studies in a manner that maximizes statistical power and improves the accuracy of inferences about cross-level interaction effects in multilevel modeling.

We should note that multilevel models come in two basic varieties, with each concerning the lack of lower level independence of observations in some high-level grouping. One form is where lower level entities are *nested* in higher level collectives (e.g., individuals in teams). Here the lower level errors are correlated by virtue of joint membership in the collective. The second form is akin to typical *repeated measures designs*, whereby some set of entities (e.g., individuals, teams, organizations) generates scores repeatedly over time. In this latter design, the lower level error terms not only are correlated but are correlated in a serial manner, given their temporal nature. The serial correlation means that the lower level unit errors are correlated in complex ways and also implies that scaling and centering of time are different than in nested arrangements, all of which present both statistical and substantive challenges (cf. Biesanz, Deeb-Sossa, Papadakis, Bollen, & Curran, 2004; Ployhart & Vandenberg, 2010). Therefore, we restrict our focus to the nested multilevel designs for this investigation. However, as we highlight below, our focus on cross-level interactions in nested multilevel arrangements still provides important guidance for designing and conducting substantive research in numerous applied psychology domains and, moreover, can serve as starting point for power estimation extensions that consider more complex, repeated measures and mixed-model designs.

Power in Multilevel Designs

As is well known, power refers to the ability of a test statistic to detect an effect of a certain magnitude with a specific degree of confidence. Generally speaking, power increases to the extent that (a) the population effect is larger; (b) sample sizes (i.e., degrees of freedom) increase; (c) the preset Type I error rate α is higher; (d) predictors and criterion are measured with higher fidelity (e.g., reliable measures, appropriate coarseness); (e) variable distributions are not restricted; and (f) assumptions of statistical tests are not violated (e.g., homogeneity of error variances, linear relationships; see Aguinis, 2004, for single-level models and Culpepper, 2010, for multilevel models). At issue is the fact that if researchers fail to have sufficient statistical power, they are susceptible to Type II errors, or the likelihood that they will falsely conclude that given effects do not exist. Such errors can lead to suboptimal use of resources, misguided interventions, deficient evaluation studies, and a wide variety of other impediments to an organization's competitive advantage and to employee welfare.

Although challenges associated with achieving sufficient power in single-level investigations are fairly well understood (see Aguinis, Beaty, Boik, & Pierce, 2005, for a review), multilevel investigations introduce additional complications. Mathieu and Chen (2011) noted that "statistical power in multilevel designs is a complex combination of the number of higher level units and lower level units under investigation, the co-variances within and between units, and a slew of other factors that are still being investigated" (p. 631). They further submitted that factors contributing to multilevel power also differ depending on the parameters of interest—namely, lower level direct effects, cross-level direct effects, or cross-level interactions. This situation has led some to simply advocate general rules of thumb for multilevel samples, such as one should have at least 30 upper level units with at least 30 lower level entities in each (i.e., the so-called 30-30 rule; e.g., Kreft & de Leeuw, 1998). However, such rules of thumb are not likely to apply universally to the wide variety of situations that

researchers encounter. Moreover, no existing approximations to date have allowed researchers to understand the impact of measurement error on multilevel power estimates. As a result, researchers cannot compute accurate a priori estimates regarding the power for their intended designs, and when they fail to find support for cross-level interactions, they cannot make informed interpretations as to whether their findings have substantive importance or might be attributable to low statistical power. In other words, we simply do not know whether hypotheses about the presence of cross-level interaction effects have been abandoned prematurely. Therefore, we turn our attention to estimating the power of tests of cross-level interactions in nested multilevel designs.

Raudenbush (1997), Raudenbush and Liu (2000), Snijders (2005), and Scherbaum and Ferreter (2009) have each highlighted some of the general issues associated with statistical power in multilevel designs. For example, consider a typical design with individual employees nested in groups. Using conventional nomenclature, the relationship between a predictor and criterion at the lower level (i.e., Level 1) of a multilevel design can be shown as

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_j) + r_{ij} \quad (1)$$

where Y_{ij} is the criterion score for the ith person in group j, β_{0j} is the intercept value for group j, β_{1j} is the slope for group j, X_{ij} is the predictor score for the ith person in group j and is centered by the group average $\bar{X}_j$, and r_{ij} is the Level 1 residual term such that $r_{ij} \sim N(0,\sigma^2)$. In most applications, the regression coefficients are assumed to be distributed jointly as random normal variables,

$$\begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} \gamma_{00} \\ \gamma_{10} \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix}\right). \quad (2)$$

That is, τ_{00} measures the variance of β_{0j}, τ_{11} measures the variance of β_{1j}, and τ_{01} measures the covariance between β_{0j} and β_{1j}. In the single predictor case, previous research notes that a way to test for the presence of cross-level interactions is to estimate the following upper level (i.e., Level 2) models (Enders & Tofighi, 2007; Hofmann & Gavin, 1998; Raudenbush, 1989a, 1989b):

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\bar{X}_j - \bar{X}) + \gamma_{02}(W_j - \bar{W}) + \gamma_{03}(\bar{X}_j - \bar{X})(W_j - \bar{W}) + u_{0j} \quad (3)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(W_j - \bar{W}) + u_{1j} \quad (4)$$

where u_{0j} and u_{1j} are residuals, or random effects, that capture group differences after controlling for Level 2 predictors $\bar{X}_j$ (group j's mean for X_{ij}), W_j (a Level 2 covariate), and the interaction between $\bar{X}_j$ and W_j. γ_{00} is the average Y_{ij} when group j is average on the other predictors (i.e., $\bar{X}_j = \bar{X}$ and $W_j = \bar{W}$) and γ_{10} is the relationship between $X_{ij} - \bar{X}_j$ and Y_{ij} when $W_j = \bar{W}$. The effect of $\bar{X}_j$ and W_j on β_{0j} is captured by γ_{01} and γ_{02}, respectively. The interaction effect between $\bar{X}_j$ and W_j is represented by γ_{03}. In Equation 4, γ_{11} captures the extent to which W_j relates to group differences in β_{1j}. The Level 2 equations can be substituted into the Level 1 equation to yield

$$Y_{ij} = \gamma_{00} + \gamma_{01}(\bar{X}_j - \bar{X}) + \gamma_{02}(W_j - \bar{W}) + \gamma_{03}(\bar{X}_j - \bar{X})(W_j - \bar{W}) + \gamma_{10}(X_{ij} - \bar{X}_j) + \gamma_{11}(X_{ij} - \bar{X}_j)(W_j - \bar{W}) + u_{0j} + u_{1j}(X_{ij} - \bar{X}_j) + r_{ij} \quad (5)$$

Equation 5 shows that the effect of the cross-level interaction between W_j and X_{ij} is captured by γ_{11}. From Equation 5, we can also see that $Var(Y_{ij}|X_{ij}) = \tau_{00} + 2\tau_{01}(X_{ij} - \bar{X}_j) + \tau_{11}(X_{ij} - \bar{X}_j)^2 + \sigma^2$ (Clarke & Wheaton, 2007).

It is important to appreciate the role of centering decisions in tests of cross-level interactions (Enders & Tofighi, 2007; Hofmann & Gavin, 1998). The observed variance in lower level predictors (X_{ij}) may reflect both lower and upper level influences, the relative extent to which can be expressed in terms of an intraclass correlation index (ICC). The ICC (ρ_x) is the ratio of between-group predictor variance (τ) relative to the total predictor variance [i.e., $\rho_x = \tau/(\tau + \sigma^2)$], where σ^2 is the variance component of the lower level residual from a null model. A null model is one where the variance of a lower level is partitioned into that which resides within and between higher level units. In this fashion, ICCs can range from 0 to 1. At issue is that if lower level predictor variance is used in its raw score form (X_{ij}) or centered on the basis of the total sample mean, ($X_{ij} - \bar{X}$), then it represents an intractable blend of upper and lower level influences. Consequently, it is not clear the extent to which an interaction involving the lower level predictor represents an interaction between the upper level moderator and the *within-group variance of the lower level predictor* versus an interaction between the upper level moderator and the *between-group variance of the lower level predictor*. For cases where one wishes to differentiate these two types of interactions, Hofmann and Gavin (1998) and Enders and Tofighi (2007) both advocated that the lower level predictor be centered within groups ($X_{ij} - \bar{X}_j$) and the between-group variance ($\bar{X}_j - \bar{X}$) be reintroduced as a Level 2 predictor. Accordingly, we implemented their recommendation.

Previous researchers have developed techniques for estimating the power of *lower level direct effects* and *cross-level direct effects* in the context of multilevel designs (cf. Raudenbush, 1997; Snijders & Bosker, 1993). Naturally, the number of lower level entities and upper level units plays a large role, akin to total sample size in single-level designs. Researchers often have control over, for example, how many individuals they can sample versus how many groups they sample. Sampling more individuals from fewer groups is usually far less costly and logistically easier than sampling a larger number of groups; however, these decisions have implications in terms of the resulting power to detect various effects. Generally speaking, there is a premium on the average number of lower level entities for enhanced power to detect Level 1 direct effects, and there is a premium on the number of upper level units for enhanced power to detect cross-level direct effects (Raudenbush & Liu, 2000). Because researchers are often interested in testing both types of effects, they are faced with a dilemma regarding how to best proceed.

Another key distinction for power in multi- versus single-level designs is the percentage of criterion variance that resides within versus between upper level units, again typically expressed in terms of an ICC. Generally speaking, lower ICCs favor the power to detect Level 1 direct effects, whereas higher ICCs favor the power to detect cross-level direct effects (Raudenbush & Liu, 2000). As is true for single-level designs, the power associated with tests in multilevel studies naturally is influenced by the magnitude of the population effect selected, the preset α level adopted, variable reliabilities and other measurement properties, variable ranges and their distributions, and the extent to which

assumptions associated with statistical tests are met.[2] Inevitably, specifying the number of parameter estimates and their complex relationships needed for conducting power analyses in multilevel designs can be a daunting task. Although the calculations associated with weighing various trade-offs are far from straightforward, available tools such as Optimal Design (Raudenbush, 1997; Spybrook, Raudenbush, Congdon, & Martinez, 2009) and Power IN Two-level designs (PINT; Bosker, Snijders, & Guldemond, 2003) make them more accessible to researchers while also permitting researchers to incorporate cost factors (Scherbaum & Ferreter, 2009). Unfortunately, Optimal Design is limited to estimating the power of treatment effects, and PINT requires users to provide estimates of a large number of parameters that may be difficult to obtain a priori (e.g., variances and covariances of residuals). Moreover, neither tool incorporates the ability to consider the impact of variable reliabilities, nor are they capable of providing power estimates for cross-level interactions. To address these needs, we provide a mechanism for addressing these shortcomings.

Our discussion thus far has highlighted that the parameters of interest in multilevel designs have direct consequences regarding the requisite information and a wide variety of decisions that researchers must consider related to the likelihood of detecting significant lower level or cross-level direct effects. However, as our review has highlighted, researchers are advancing hypotheses and testing *cross-level interactions* (i.e., testing varying slope models) at a rapidly growing rate. Whereas there are formulae and tools available for multilevel designs with fixed direct effects, Scherbaum and Ferreter (2009) 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" (p. 363). Indeed, Snijders and Bosker (1993) noted the following: "In the more general case, with within-group regressions which may be heterogeneous, it is impossible to derive analytical expressions for optimal sample sizes" (p. 249). Over a decade later, Snijders (2005) echoed the sentiment 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" (p. 1572). The intractability of these tests stem from the fact that they *simultaneously* hinge on factors associated with the estimation of the lower level slopes and the higher level moderation of them. Notably, Zhang and Willson (2006) used simulation data to investigate the power difference among three multilevel analytic techniques: (a) hierarchical linear modeling (HLM); (b) deviation structural equation models; and (c) a hybrid approach of HLM and structural equation models. Although their study yielded some important insights concerning different multilevel analytic techniques, it did not offer a usable general approach for estimating the power of cross-level interactions. Zhang and Willson also considered fewer parameters and values per parameter than we do here—for example, they did not consider the role of measurement error.

Accordingly, we next present the results from a simulation study in which we manipulated research design and measurement factors that previous research has suggested could influence the power to detect cross-level interactions. In the simulation study we examined the statistical power of cross-level interaction tests and explored the relative impact of various factors on statistical power.

Simulation Study

Data Generation Procedure

We conducted a Monte Carlo simulation to understand how research design and measurement features affect the statistical power to detect cross-level interactions in multilevel models. For the simulation study, we employed the conventional alpha level of .05 for consistency and comparability purposes. Naturally, one's preselected alpha level is directly related to the power of any statistical test, so, as we describe later, a power calculator that we created and make available for use allows researchers to specify any desired alpha level.

Let Y_{ij} and X_{ij} be true scores for the Level 1 criterion and predictor, respectively, for the *i*th person in the *j*th Level 2 unit. Also, let W_j be a Level 2 predictor assessed without measurement error. The power to detect the cross-level interaction in Equation 5 was estimated by simulating X_{ij}, Y_{ij}, and W_j as random variables from normal distributions. Additionally, $X_{ij} \sim N(\mu_j, \sigma^2_\mu + \sigma^2_X)$ where σ^2_X is the within-unit variance of X_{ij}, which was standardized within each unit, that is,

$$\sigma^2_\mu = \frac{\rho_X \sigma^2_X}{1 - \rho_X} = \rho_X \sigma^2_X (1 - \rho_X)^{-1}.$$

The variability across Level 2 units (σ^2_μ) was allowed to vary based upon the ICC for X_{ij}, ρ_x. Specifically,

$$\rho_X = \frac{\sigma^2_\mu}{\sigma^2_\mu + \sigma^2_X} = \sigma^2_\mu (\sigma^2_\mu + \sigma^2_X)^{-1}.$$

The dependent variable, Y_{ij}, was generated using the following equation:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_j) + r_{ij}\sqrt{1 - \beta^2_{1j}} \quad (6)$$

where β_{0j} and β_{1j} are a random intercept and slope, respectively, and the within-unit error term $r_{ij} \sim N(0,1)$. Moreover, β_{1j} represents the correlation between X_{ij} and Y_{ij} within each Level 2 unit *j*. The random coefficients were generated with the following equations:

$$\beta_{0j} = \sigma_j\left(\gamma_{0X}\frac{\bar{X}_j}{\sigma_\mu} + \gamma_{0w}W_j + \gamma_{0Xw}\frac{\bar{X}_j W_j}{\sigma_{Xw}} + u_{0j}\right) \quad (7)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{1w}\sigma_{\beta_1}W_j + \mu_{1j}. \quad (8)$$

W_j was standardized with a mean of zero and unit variance, and the Level 2 error terms, u_{0j} and u_{1j}, were generated with $u_{0j} \sim N(0,\tau_{00})$ and $u_{1j} \sim N(0,\tau_{11})$. Also, note that X_j and W_j were simulated as uncorrelated random variables.

The between-group variance for Y_{ij} (σ^2_j) was manipulated by changing the ICC value, ρ_y. That is, $\rho_y = \sigma^2_j/(\sigma^2_j + \sigma^2)$. In this study, $E(X_{ij}) = 0$, because $E(\mu_j) = 0$, and the random effects were

[2] Other factors play prominent roles in multilevel power estimates, including the specific estimation method that is employed, such as restricted maximum likelihood (RML) versus full maximum likelihood estimation (FML) versus Bayesian approaches (see Raudenbush & Bryk, 2002; Scherbaum & Ferreter, 2009), and the presence of covariates (i.e., Level 1 or 2; see Raudenbush, 1997). For our purposes, we adopt RML as implemented in the popular hierarchical linear modeling software (HLM; see Raudenbush & Bryk, 2002; Raudenbush et al., 2004) and exclude covariates.

assumed to be independent (i.e., $\tau_{01} = 0$),[3] so $\sigma_j^2 = \sigma^2\rho_y/(1 - \rho_y)$ for the average X_{ij} (Clarke & Wheaton, 2007; Maas & Hox, 2004) and τ_{00} is the variance of β_{0j} that is independent of $\bar{X}_j$, W_j, and $\bar{X}_jW_j$. Specifically, $\tau_{00} = \sigma_j^2(1-\gamma_{0\bar{X}}^2 - \gamma_{0w}^2-\gamma_{0\bar{X}w}^2)$. In Equation 7, $\gamma_{0\bar{X}}$, γ_{0w}, and $\gamma_{0\bar{X}w}$ capture the standardized relationships between $\bar{X}_j$, W_j, and $\bar{X}_jW_j$ with β_{0j} Additionally, γ_{1w} is the standardized relationship between W_j and β_{1j} and γ_{10} is the overall average correlation between X_{ij} and Y_{ij} across Level 2 units. In Equation 8, the standard deviation of slopes is $\sqrt{\tau_{11}}$ and the unique variance in slopes (i.e., β_{1j}), after controlling for W_j, is $\tau_{11} = \sigma_{\beta_1}^2(1-\gamma_{1w}^2)$. It is important to emphasize that variability in intercepts and slopes was introduced by manipulating σ_j^2 and $\sigma_{\beta_1}^2$, whereas τ_{00} and τ_{11} represent the conditional variance of β_{0j} and β_{1j} after controlling for Level 2 predictors.

In our simulation we also examined the effect of measurement error on the power of the cross-level interaction test. In particular, measurement error was introduced into the true scores of X_{ij}, Y_{ij}, and W_j to create fallible measures (x_{ij}, y_{ij}, and w_j) using the following equations:

$$x_{ij} = \sqrt{\rho_{xx}}\, X_{ij} + \sqrt{1 - \rho_{xx}}\, e_{xij} \quad (9)$$

$$y_{ij} = \sqrt{\rho_{yy}}\, Y_{ij} + \sqrt{1 - \rho_{yy}}\, e_{yij} \quad (10)$$

$$w_j = \sqrt{\rho_{ww}}\, W_j + \sqrt{1 - \rho_{ww}}\, e_{wj} \quad (11)$$

where e_{xij}, e_{yij}, and e_{wj} are standard normal error terms and ρ_{xx}, ρ_{yy}, and ρ_{ww} are reliability coefficients of x_{ij}, y_{ij}, and w_j, respectively.

Simulation Design

Table 2 summarizes the manipulated parameters and the values that we used in the simulation study. We selected parameter values to be representative of the extant literature as well as to be those encountered by most applied psychology and management researchers. For example, on the basis of our review of 79 multilevel investigations published in the *Journal of Applied Psychology* between 2000 and 2010, the average Level 1 unit sample sizes ranged from 2 to 291, with a median of 5, whereas the Level 2 sample sizes ranged from 12 to 708, with a median of 51. A few recent atypically large studies (e.g., Atwater, Wang, Smither, & Fleenor, 2009; Dierdorff, Rubin, & Morgeson, 2009) markedly skewed these distributions, as their 85th percentiles for Level 1 sample sizes was about 18, whereas the 85th percentile of the Level 2 sample sizes was about 51. The studies that we feature in Table 1 are representative of the ones published in the *Journal of Applied Psychology* over this period, although Mathieu et al. (2007) had both relatively small Level 1 samples sizes and a relatively large Level 2 sample. Accordingly, in our simulation we set the average Level 1 N (n_i) to represent the primary range of studies (i.e., 3, 5, and 7) and we included a condition with 18 to capture the upper portion of the distribution. We varied the Level 2 N (n_j) across 20, 40, and 60 and also included 115 to capture the corresponding portion of its distribution.

Table 2
Parameters and Parameter Values Used in Monte Carlo Simulation

Symbol	Parameter	Values
n_i	Average L1 N	3, 5, 7, 18
n_j	L2 N	20, 40, 60, 115
ρ_x	L1 X ICC	.15, .30
ρ_y	L1 Y ICC	.15, .30
ρ_{xx}	L1 X reliability	.8, .9, 1.0
ρ_{yy}	L1 Y reliability	.8, .9, 1.0
ρ_{ww}	L2 W reliability	.8, .9, 1.0
γ_{10}	Average L1 direct effect	0, .2, .4
$\sqrt{\tau_{11}}$	*SD* of L1 slopes	.10, .17, .22
γ_{1w}	Cross-level interaction	0, .15, .3, .45, .75
$\gamma_{0\bar{x}}$	Direct effect of $\bar{X}_j$	0, .1, .2
$\gamma_{0\bar{x}w}$	Direct effect of $\bar{X}_jW_j$	0, .1, .2
γ_{0w}	Cross-level direct effect of W_j	0, .15, .3, .45
Total cells		2,799,360
No. replications		1,000
Total no. score sets		2,799,360,000

Note. L1 = Level 1; L2 = Level 2; ICC = intraclass correlation coefficient; *SD* = standard deviation.

We set variable ICCs (ρ_x and ρ_y) to range from .15 to .30, which correspond to fairly moderate and large values (Maas & Hox, 2005; Scherbaum & Ferreter, 2009) and capture the range (i.e., .00 to .39) observed in Table 1. We set variable reliabilities (ρ_{xx}, ρ_{yy}, and ρ_{ww}) at .8, .9, and 1.0, which are also representative of the values reported in Table 1. The remaining parameter values are among the most difficult to establish because they are not reported consistently (if at all) in the literature. Indeed, the primary guidance noted in previous literature stems from recommendations offered by Raudenbush and Liu (2000) who, in turn, simply applied Cohen's (1988) rules of thumb for small, medium, and large effect sizes. Therefore, we obtained raw data and calculated these directly for each of the studies reported in Table 1.[4] From a Level 1 model that included only the lower level predictor, the average slopes (γ_{10}) ranged from −.06 to .45 with a mean of .17 across the seven analyses. Accordingly, we varied the magnitude of the average lower level slopes from 0 to .2 to .4.

Given the small sample sizes and the fact that Level 1 sample size distributions were markedly skewed, we conducted meta-analyses to estimate the variability (i.e., *SD*) of slopes $\sqrt{\tau_{11}}$. In particular, we applied Hunter and Schmidt's (2004) "bare-bones" meta-analysis, which corrects only for varying study (in this case, Level 1) sample sizes. As illustrated in Table 1, the meta-analysis-based *SD* of slopes ranged from .00 in one of the Mathieu et al. (2007) analyses to .27 in Hofmann et al. (2003), averaging .15. Based on these findings and those of the individual study meta-analyses, we set the *SD*s of Level 1 slopes at .10, .17, and .22. Finally, we set the magnitudes of the cross-level direct effect (γ_{0w}) to 0, .15, .30, and .45 and the cross-level interactions (γ_{1w}) to 0, .15, .30, .45, and .75, which are consistent with the ranges reported in Table 1 and recommendations from the previous work cited above. In total, our simulation design resulted in 2,799,360 unique conditions (i.e., combinations of parameter values). Because we replicated each set of parameter values 1,000 times, our simulation generated and used almost 2.8 billion sets of scores.

[3] In support of this assumption, analyses of the data from studies included in Table 1 revealed that τ_{01} absolute values ranged from .0002 to .06, averaging .02.

[4] We thank the study authors for graciously supplying us with their raw data. More detailed information regarding these additional analyses and results are not reported in this paper but are available from the first author.

Simulation Accuracy Checks

Prior to conducting the substantive analysis, we performed a detailed investigation to verify the accuracy of the simulation procedures. We examined the Type I error rate of cross-level interactions and bias of parameter estimates in cases where artifacts such as measurement error were absent from the data generation procedure. For the simulation, we obtained 1,000 replications for each parameter value combination in the absence of a cross-level interaction (i.e., $\gamma_{1w} = 0$). We conducted key accuracy checks for Type I error rates using the 559,872 cells from our design that included a true null cross-level interaction effect.

The average empirical Type I error was .045, and the median absolute value of the difference between the empirical Type I error rates and the expected (i.e., prespecified Type I error) rate of .05 across all Type I error rate conditions was .007. As a further check we found that 95% of these values fell between .000 and .022. The similarity of the empirical Type I error rates compared to the a priori set Type I error rate is evidence to support the accuracy of the data-generation procedures. In short, the results confirmed the ability of the generation procedure to provide data suitable for substantive analyses.

As a second and complementary set of accuracy checks, we investigated the extent to which the simulation generated cross-level interaction effects that were unbiased in conditions without statistical artifacts. It is important to conduct these analyses regarding the potential presence of bias for the conditions in the simulation for which methodological and statistical artifacts are absent, given their known biasing effects on the observed effect sizes. In this second set of key accuracy checks, we isolated the 2,239,488 cells from our simulation for which there was a true population cross-level interaction effect (i.e., $\gamma_{1w} > 0$). The absolute-value differences between the empirical (i.e., Monte Carlo generated) and the population values (i.e., prespecified parameter values) were negligible, with a median absolute of .007. Moreover, 95% of the absolute-valued differences fell between .000 and .032 for γ_{11}. Thus, the simulation successfully modeled the cross-level relationships as anticipated. Given the evidence in support of the validity of our simulation, we next describe the simulation's substantive results.

Results

Our goal in the simulation study was to examine factors that affect power in cases where the cross-level interaction (i.e., γ_{1w}) is greater than zero. The average statistical power across the conditions in the simulation for detecting a cross-level interaction was only .192. We investigated the relative importance of the various factors we manipulated by conducting an ANOVA with statistical power as the dependent variable. The different levels of the factors accounted for 96.1% of the variation in statistical power and are summarized in Table 3. The partial eta-squared percentage (P-η^2%) values in Table 3 suggest that the following four main effects, in order of importance, were the primary factors in determining the statistical power of cross-level interaction tests: (a) size of the cross-level interaction (γ_{1w} P-η^2% = 33.08, $p < .001$); (b) average Level 1 sample size (n_i P-η^2% = 18.18, $p < .001$); (c) Level 2 sample size (n_j P-η^2% = 11.46, $p < .001$); and (d) the standard deviation of slopes $\sqrt{\tau_{11}}$ (P-η^2% = 9.31, $p < .001$). With three exceptions, the remaining direct effects were all statistically significant ($p < .001$) but had negligible effects. The ICC effects for X_{ij} and Y_{ij} (i.e., ρ_y and ρ_x) and reliabilities of x_{ij}, y_{ij}, and w_j were statistically significant but explained small percentages in the variance of statistical power values (<1%). The insignificant effect of ρ_y and ρ_x is theoretically grounded. For instance, an anonymous reviewer noted that group-mean centering X_{ij} removes Level 2 variance and the relationship between X_{ij} and Y_{ij} reflects an association between within group variances. Consequently, group-mean centering removes Level 2 variance, and ρ_y and ρ_x do not affect the power of detecting the cross-level interaction. The three statistically nonsignificant direct effects were (a) Level 2 moderator direct effect, W (γ_{0w}); (b) average Level 1 predictor reintroduced at Level 2 ($\gamma_{0\bar{x}}$); and (c) average Level 1 predictor by Level 2 interaction, computed at Level 2 ($\gamma_{0\bar{x}w}$).

Table 3
Results of Analysis of Variance of the Effects of Manipulated Parameters on the Statistical Power to Detect a Cross-Level Interaction Effect

Effect	Effect description	P-η^2%
Main effects		
n_i	Average L1 N	18.1754
n_j	L2 N	11.4557
γ_{1w}	Cross-level interaction	33.0788
γ_{10}	Average L1 direct effect	0.0513
$\sqrt{\tau_{11}}$	*SD* of L1 slopes	9.3129
ρ_x	L1 X ICC	.00002
ρ_y	L1 Y ICC	.00008
ρ_{xx}	L1 X reliability	0.2259
ρ_{yy}	L1 Y reliability	0.2434
ρ_{ww}	L2 W reliability	0.2258
γ_{0w}	Cross-level direct effect, W	.00000
$\gamma_{0\bar{x}}$	Direct effect of $\bar{X}_j$	.00000
$\gamma_{0\bar{x}w}$	Direct effect of $\bar{X}_j W_j$	.00001
Interaction effects		
$n_i * n_j$		9.4071
$n_i * \gamma_{1w}$		7.1435
$n_i * \sqrt{\tau_{11}}$		1.0217
$n_j * \gamma_{1w}$		7.2169
$n_j * \sqrt{\tau_{11}}$		1.8559
$\gamma_{1w} * \sqrt{\tau_{11}}$		5.1652

Note. The model accounted for 96.1% of the variance in statistical power values. All effects are statistically significant at $p < .001$ except for γ_{0w}, $\gamma_{0\bar{x}}$, and $\gamma_{0\bar{x}w}$ ($p < .05$). P-η^2% = partial eta-squared percentage; L1 = Level 1; L2 = Level 2; ICC = intraclass correlation coefficient; *SD* = standard deviation.

Beyond the main effects, there was also evidence that six two-way interactions affected the power to detect cross-level interactions, although the relative influences of these effects were smaller than the aforementioned main effects. The significant interaction between the Level 2 sample size and the magnitude of the cross-level interaction ($n_j * \gamma_{1w}$ P-η^2% = 7.14, $p < .001$) is presented in Figure 1. As shown in this figure, power remains under 50% for all combinations until one approaches the highest Level 2 sample size (i.e., 115) with relatively large cross-level interactions (i.e., $\gamma_{1w} = .75$). The average Level 1 sample size also interacted with the magnitude of the cross-level interaction ($n_i * \gamma_{1w}$ P-η^2 = .071, $p < .001$), as shown in Figure 2. Here again, power remains < 50% until Level 1 average sample size gets to be

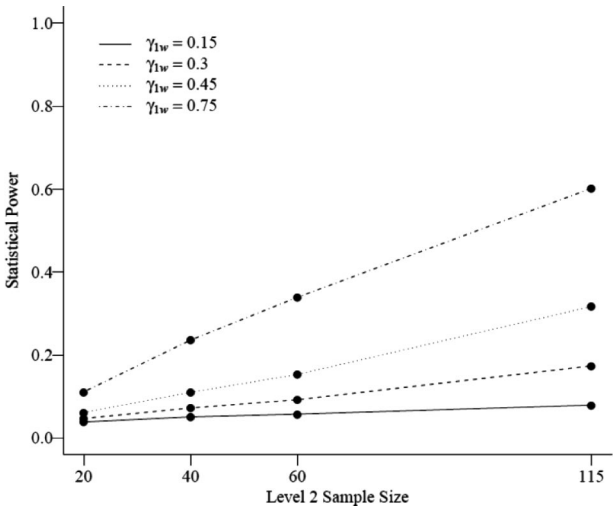

Figure 1. Statistical power to detect cross-level interactions as a function of Level 2 sample sizes and the magnitude of the cross-level interaction effect (γ_{1w}).

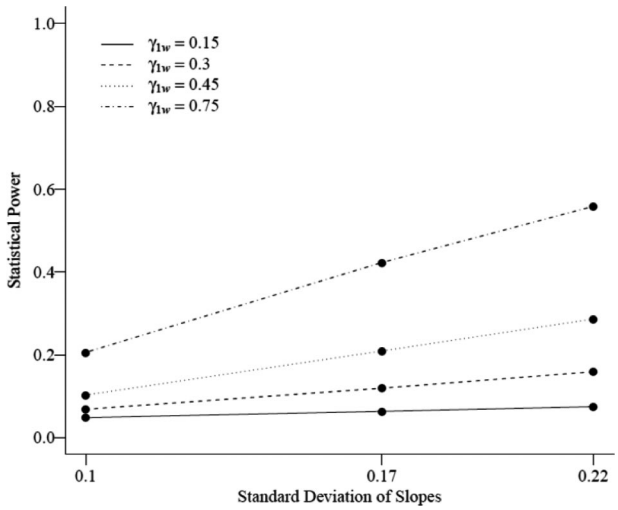

Figure 3. Statistical power to detect cross-level interactions as a function of the standard deviation of slopes and the magnitude of the cross-level interaction effect (γ_{1w}).

around 10, and even then it must be paired with relatively large cross-level interactions (i.e., $\gamma_{1w} = .75$). With Level 1 samples sizes around 18, the large cross-level interactions reach 90% power, whereas moderate sized cross-level interactions (i.e., $\gamma_{1w} = .45$) had power around 60%.

The standard deviation of slopes interacted significantly with the magnitude of the cross-level interaction ($\sqrt{\tau_{11}} * \gamma_{1w}$ P-η^2% = .052, $p < .001$), Level 2 sample size ($\sqrt{\tau_{11}} * n_j$ P-η^2% = .019, $p < .001$), and average Level 1 sample size ($\sqrt{\tau_{11}} * n_i$ P-η^2% = .010, $p < .001$), as depicted in Figures 3, 4, and 5, respectively. Figure 3 shows that power remains less than 50% for the range of slope and cross-level interaction magnitude values that we examined. Extrapolating from the findings in Figure 3, statistical power would not reach 80% for even a relatively large cross-level interaction effect (i.e., $\gamma_{1w} = .75$) until the *SD* of slopes exceeded .3, which is beyond typically observed values. Similarly, power remains less than 50% for the combinations of *SD* of slopes and Level 2 samples size values that we examined. Extrapolating the findings in Figure 4 indicates that power would reach 80% for relatively large Level 2 samples (i.e., 115) and *SD* of slope of approximately .5. The power associated with combinations of the *SD* of slopes and the average Level 1 sample sizes (see Figure 5) showed a similar pattern as that for Level 2 sample size but accelerated at a faster pace. In other words, for relatively large average Level 1 sample sizes (i.e., 18), power exceeded 50% for *SD* of slopes at .17, although it would not reach 80% unless the *SD* of slopes exceeded .3. Note that we performed these extrapolations

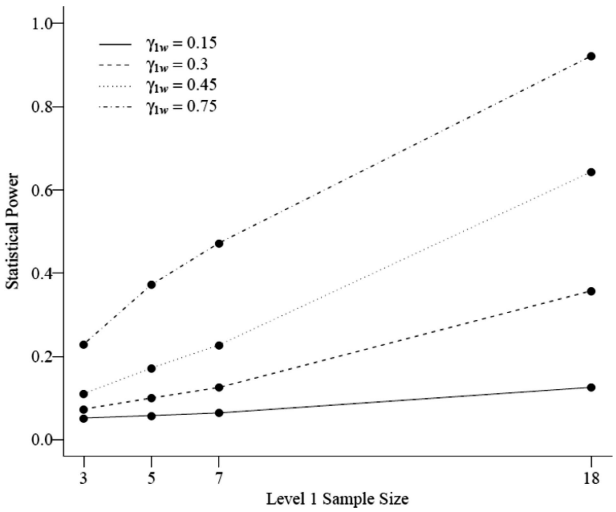

Figure 2. Statistical power to detect cross-level interactions as a function of average Level 1 sample sizes and the magnitude of the cross-level interaction effect (γ_{1w}).

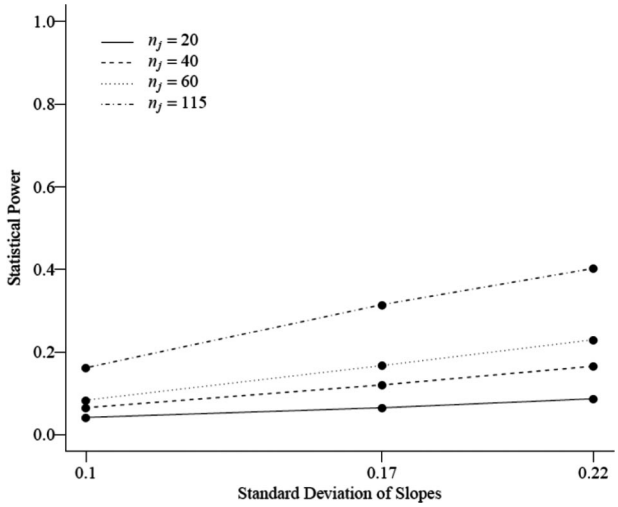

Figure 4. Statistical power to detect cross-level interactions as a function of the standard deviation of slopes and Level 2 sample size (n_j).

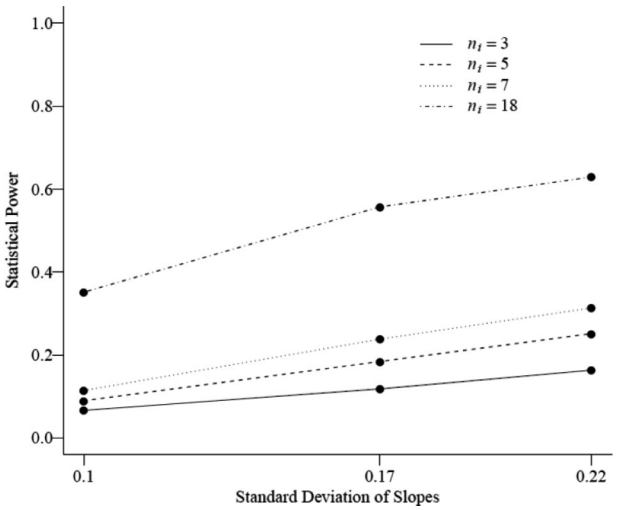

Figure 5. Statistical power to detect cross-level interactions as a function of the standard deviation of slopes and the average of Level 1 sample sizes (n_i).

using predictions from polynomial regressions derived from the curves depicted in Figures 3–5, and they result in values that are not likely to occur in practice.

Finally, the two samples sizes interacted significantly ($n_i * n_j$ P-η^2% = .009, $p < .001$), as illustrated in Figure 6. As shown, statistical power remains very low (i.e., < 40%) for the smaller average Level 1 samples sizes (i.e., ≤ 7) even if paired with a relatively large Level 2 sample size of 115. However, relatively large average Level 1 samples sizes (i.e., ≥ 18) afford power > 60% with Level 2 samples as small as 25 and surpass power of 80% with Level 2 samples of 35. Also, using these findings we investigated power values associated with the popular 30-30 rule (Kreft & de Leeuw, 1998), given its predominance in the literature. Holding both Level 1 and Level 2 samples sizes at 30 and averaging across other parameter values in our entire simulation, the 30-30 rule yields statistical power values ranging from 6.5% for relatively small cross-level interactions (i.e., $\gamma_{1w} = .15$) to 79.5% for relatively large cross-level interactions (i.e., $\gamma_{1w} = .75$), averaging 32.4% across all other parameter values.

In sum, the results of the simulation study illustrate how important research design and measurement features relate to the power to detect cross-level interactions. Moreover, results indicate that the power to detect cross-level interactions depends, in large part, on the magnitude of the cross-level interactions and the *SD* of slopes, both in and of themselves but also in combination. The Level 2 and average Level 1 sample sizes also play prominent roles in determining the ability to detect cross-level interactions, both directly and in combinations with other factors. Importantly, our simulation results also demonstrate the fallacy of relying on simplistic rules of thumb, such as the 30-30 rule.

Discussion

Accurate interpretations about the presence of cross-level interactions lie at the heart of multilevel investigations in applied psychology. Our goal was to assess the factors that affect statistical power to detect cross-level interaction effects and to address Snijders' (2005) observation that there is not clear formula for some general cases of complex multilevel models. Accordingly, we sought to provide researchers with a means to a priori estimate the power of their cross-level interaction tests. Our simulation study led us to conclude that our power approximation procedure provides an accurate method to calculate the power of tests of cross-level interactions. Substantively, key insights include the fact that the power associated with such tests is affected primarily by the magnitude of the direct cross-level effect (γ_{1w}) and the standard deviation of the Level 1 slope coefficients ($\sqrt{\tau_{11}}$), as well as by both average lower level (n_i) and the upper level (n_j) sample sizes. Moreover, the same four factors worked in combination to drive power estimates, suggesting that there can be compensatory factors affecting the power to detect cross-level interactions.

The lion's share of the variance in power values clearly was accounted for by the magnitude of the cross-level interaction effect, both as a main effect and in combination with the Level 2 and average Level 1 sample sizes. This is consistent with research regarding statistical power in other data-analytic contexts, in that the magnitude of the effect being investigated plays a prominent role, with larger effects affording greater power. In the cross-level context, however, the variability (i.e., *SD*) of the Level 1 slopes also plays a significant role, both as a main effect and in combination with samples sizes and the magnitude of the cross-level effect. In short, testing for relatively large cross-level interactions in instances where there is abundant variability of Level 1 slopes provides the greatest opportunity to unearth significant effects.

Naturally, the Level 2 and average Level 1 sample sizes are focal points for cross-level studies. These represent the sampling frames for investigators and drive decisions as to how to best allocate resources. They are also the factors that are more likely to be under researchers' control. However, contrary to conventional wisdom, our findings suggest that the average Level 1 sample size has a relative premium of about 3:2 as compared to the Level 2 sample size (cf. Snijders, 2005). Notably, both levels' samples sizes interacted significantly with the magnitude of the cross-level interaction and with the variability of the Level 1 slopes, with the

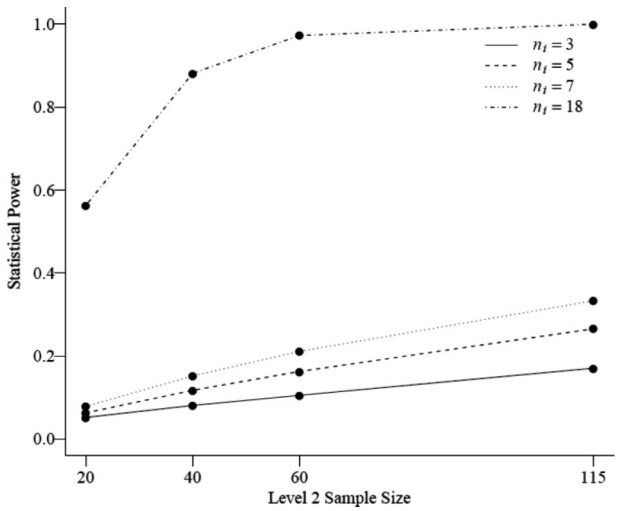

Figure 6. Statistical power to detect cross-level interactions as a function of Level 2 sample sizes and the average of Level 1 sample sizes (n_i).

corresponding effect sizes being comparable in all instances. Importantly, we scaled the two sample size distributions to be comparable and representative of cross-level studies published in the *Journal of Applied Psychology* from 2000 to 2010 (Cooper & Richardson, 1986). These findings suggest that researchers interested in testing cross-level interactions should attempt to sample more thoroughly within units to enhance their power, as compared to sampling a more units. Yet, many times in applied psychology there may be a natural limit to the sizes of certain groups (e.g., work teams, classrooms, nuclear families), which may preclude large numbers. Fortunately, our results suggest that in such instances some additional power may be garnered by sampling more Level 2 units. 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. In other words, if 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.

Multilevel (ML) Power Tool

We developed a Monte Carlo program (see Appendix), executable in R (R Development Core Team, 2011), for researchers to a priori estimate the power of their cross-level interaction tests. R is a language and environment for statistical computing and graphics that can be run locally on a personal computer or via Web interfaces (see, e.g., Culpepper & Aguinis, 2011, for a review of R). In order to a priori calculate the power of a cross-level interaction test, one needs to estimate values for the various parameters included in Equation 5 and listed in Table 2. An ideal situation would be to conduct a pilot study with a representative sample from one's target population and estimate values directly from those data. For purposes of consistency in these power calculations, we advocate standardizing (i.e., *Z* scoring) all lower level variables on the basis of the total sample to provide a common metric (for similar recommendations, see Hox, 2010, p. 243; Raudenbush & Liu, 2000, p. 202). These scores can then be used to derive the within-group centering and average Level 1 predictor scores for reintroduction at Level 2, as outlined earlier (see Enders & Tofighi, 2007; Hofmann & Gavin, 1998). We also suggest standardizing all Level 2 effects at that level. Although researchers often do not have the time or resources to conduct elaborate pilot studies, in instances where the primary study will represent an expensive endeavor or when it is vital to have high confidence that one is avoiding Type II errors, they may well be justified.

If a pilot study cannot be conducted, several parameter values may be available from previous investigations. For example, previous work and perhaps even meta-analyses may provide insights regarding the magnitude of lower level and cross-level direct effects. Our experience has been that although researchers sometimes report ICCs for their criteria, they rarely do so for their Level 1 predictors. In lieu of such information, researchers might consider adopting values ranging from very small (e.g., .02) to fairly high (e.g., .75) for the ICC of lower level predictors, as we did earlier in our simulation.

Values for the variability of slopes and the magnitude of the cross-level interaction are not likely to be available from previous research and are among the most difficult to estimate. Generally speaking, the variability of slopes and the magnitude of cross-level interactions are likely to be higher in instances when disordinal (i.e., crossover) interactions are anticipated, as compared to when ordinal (i.e., noncrossing) interactions are expected. In the absence of information, however, we recommend considering a wide range of values. Note that the magnitudes of these parameters are influenced by standardization and centering techniques that are employed, as well as by substantive factors. For example, some have advocated conventionally small (.20), moderate (.30), and large (.40) average effect sizes (e.g., Raudenbush & Liu, 2000; Snijders & Bosker, 1993, 1999), whereas the range of values observed in the studies we chronicled in Table 1 was from .00 to .27. We observed markedly variable and skewed distributions of Level 1 slopes in the studies we reviewed in Table 1, but bare bones meta-analysis revealed values of .10, .17, .and 22 as representing relatively little, moderate, and high variability, respectively. Little is known about the true distributions of slopes, and we strongly advocate that researchers leverage prior work in their substantive area and/or conduct a pilot study to choose reasonable values for their applications. In many instances, values (or ranges) for various parameters should be chosen for their substantive importance. For example, effect sizes of certain magnitudes may be associated with strategic imperatives or goals, break-even analyses, valued impact of some social initiative, or points where an effort becomes cost prohibitive.

To demonstrate the application of ML Power Tool, as well as to illustrate the statistical power of previously published tests of cross-level interactions, we calculated the parameter values directly from the raw data for the studies listed in Table 1. We inserted those values in the syntax that appears in the Appendix and then entered and executed the program code in R. We emphasize that these calculations are for illustrative purposes and that proper power estimation should be done in an a priori, not post hoc, fashion. That said,

> Post hoc power analysis is not only useful in evaluating one's own analysis, as just shown, but also in the planning stages of a new study. By investigating the power of earlier studies, we find what effect sizes and intraclass correlations we may expect, which should help us to design our own study. (Hox, 2010, p. 240)

As shown in the rightmost column of Table 1, the power to detect cross-level interactions across three levels of alpha (i.e., .10, .05, and .01) was quite low in previous research. At the $\alpha = .05$ level, the estimates ranged from .19 in Liao and Rupp (2005) to .80 in Hofmann et al. (2003), with an average of .40. At the $\alpha = .01$ level, the range was .06 to .59 ($M = .22$), whereas at the $\alpha = .10$ level the estimates ranged from .29 to .88 ($M = .51$). In short, the power to detect significant cross-level interactions, even under fairly advantageous conditions, is quite low and substantially below the conventional .80 level. Of course, finding statistically significant effects in spite of low power can also be indicative of the strength of the true population effects considered in these previously published studies.

Notably, several of these authors commented on the power of their cross-level tests. For example, Liao and Rupp (2005) mentioned that "our relatively small number of groups constrained our statistical power to detect the hypothesized [cross-level] relationships" (p. 253). However, they had no way of knowing just how

limited the power to detect significant interactions was in their study. Elsewhere, focusing on their Level 2 sample size ($N = 212$) and the fact that their cross-level interaction hypotheses were supported, Mathieu et al. (2007) mistakenly asserted that they had "sufficient power to test our hypotheses adequately" (p. 536). We now know that Mathieu et al. had less than a one in three chance at finding significant ($p < .05$) cross-level interactions in their data. Having a method by which one can determine the power to detect significant cross-level interactions clearly will sharpen one's ability to draw more accurate conclusions from multilevel investigations.

Implications

Application of our power calculator to previous research suggests that, generally speaking, the power to detect significant cross-level interactions is quite modest. One of the primary implications of this work is that researchers should exercise caution when interpreting statistically nonsignificant cross-level interaction tests. Although the limited power of single-level tests of interactions has been lamented for quite some time, it appears that the power to detect cross-level interactions is just as low, if not even lower. Naturally, a second direct implication of this work is that researchers should calculate their anticipated power a priori and employ the largest samples possible. Whereas using reliable measures certainly enhances power, beyond a certain point (e.g., .70) the impact of improved reliability was negligible as compared to that afforded by larger samples, at both lower and upper levels. We hasten to add that we are not suggesting that measurement reliability has no implications for statistical power—indeed, it is very relevant. Although the reliability of measures had relatively little impact on power in our simulation, that was largely attributable to the fact that we used distributions of reliabilities from articles published in the *Journal of Applied Psychology*. Of course, given that it is a premier outlet for applied psychological research, studies with poor measurement stand little chance of appearing in the journal. In effect this amounts to reliability distributions that may be restricted as compared to the population of cross-level investigations. That is not, however, likely to be the case for the other parameters on our simulation.[5]

By way of example, assume that the reliabilities of the lower level predictor and upper level moderator are both .60 and their latent correlation is $r_{xw} = .30$. Applying Bohrnstedt and Marwell's (1978) formula (designed to estimate single-level product reliabilities) to these values as a rough approximation would suggest that the product term reliability would be about .41! If the predictor and moderator reliabilities were .90, then the reliability of the product term would be closer to .83. The implication of this difference in reliabilities is that the observed cross-level effect sizes are roughly 90% of the magnitude of true (corrected for attenuation) effect sizes with reliabilities around .90, whereas the observed cross-level effects sizes will be only about 65% of the true effect sizes if estimated with measures having reliabilities around .60. We derived those values using traditional corrections for attenuation, and they hold throughout the range of effect sizes. We wish to emphasize that these estimates are exceedingly rough guesses, as the proper measurement model for cross-level interactions has yet to be specified. With that caveat, we surmise that researchers using measures with poor versus high reliabilities (say .60 vs. .90) may be reducing the power to detect significant cross-level interactions by around 25%. In sum, although measurement reliability played a limited role in our simulation, it is very important for the power associated with testing cross-level interactions. This clearly represents an important direction for future research and development. But equally clear is that measurement reliability is at a premium for researchers interested in testing cross-level interactions.

When it comes to the power of cross-level interaction tests, our findings suggest that there is about a 3:2 premium on the average size of the lower level samples, as compared to the upper level sample size. In other words, researchers wanting to conduct accurate tests of cross-level interactions should place relatively more emphasis on sampling larger units, as compared to sampling a larger number of units. These empirically based conclusions differ from those offered by Snijders and Bosker (1993) and others summarized in Scherbaum and Ferreter (2009), who placed more emphasis on sampling upper level units. Whereas more research is certainly warranted on the topic, our simulation was fairly comprehensive, manipulated 13 different parameters with values derived from the extant literature, and included a large number of replications.

Our results also highlight the importance of sampling frames for multilevel investigations. Gaining access to a sufficiently large number of units, whether they are teams, departments, or organizations, is an expensive and time-consuming task. Quite often researchers gain access through a single organization that has branches, stores (e.g., Chen et al., 2007), sales regions (e.g., Mathieu et al., 2007), or teams (e.g., Mathieu & Taylor, 2007) that perform comparable work under similar circumstances. Elsewhere, researchers who employ large-scale secondary sources often limit their investigations to, for example, organizations that fall within certain standard industrial classification (SIC) codes. These types of sampling strategies certainly facilitate collecting data for a multilevel study while holding a slew of potential contaminating influences constant, but they also limit the degree to which lower and upper level relationships are likely to vary within and across levels.

Take, for example, an investigation in which one believes that an upper level variable (e.g., team cohesion) moderates a lower level relationship (e.g., individuals' job satisfaction–withdrawal behaviors relations). Let us further assume that organizations have fairly strong cultures and that attraction–selection–attrition processes (Schneider, Goldstein, & Smith, 1995) operate to create fairly comparable teams, each with fairly homogeneous members. To the extent that teams are sampled from a relatively smaller, as compared to greater, number of organizations (i.e., in effect, a third-level variable), the variability of the job satisfaction–withdrawal slopes is likely to be diminished, as will be the magnitude of the cross-level interaction effect. Both factors would

[5] To further explore the influence of reliability in this context, we ran a second simulation and expanded the range of all three reliabilities from 0.8–1.0 to .6–1.0 while limiting the ranges for the other parameters to their most representative values. In other words, we substantially favored the odds of reliabilities playing a larger role in the overall power estimates. Whereas the combined linear effects of measurement reliability were <1% in our original study, they were still < 5% in this second simulation, even when biasing it to favor their influence. Details regarding this supplemental simulation are available from the authors.

serve to seriously limit the power to detect a significant cross-level interaction. In short, there is a premium not only on sampling a larger number of teams in this case but also on *sampling a greater diversity of teams* from a greater number of organizations or settings so as to not constrain variance across levels. Notably, this suggestion harkens to similar concerns with "extreme-groups designs" whereby researchers intentionally sample entities from the ends of a predictor distribution to maximize variance in a sample. So doing raises questions about the representativeness of the resulting effect sizes because they are subject to range enhancement (see Aguinis, 2004; Cortina & DeShon, 1998). However, our recommendation to sample a greater diversity of settings is not intended to artificially enhance the range of predictor values but rather to better represent the target population of entities. In other words, researchers are typically interested in making generalizations to groups, units, organizations, or other collectives as a whole, rather than to those that reside in a single organization or industry sampled in a typical study. In sum, the larger issue is that researchers working with multilevel designs should pay serious attention to sampling issues associated with both lower and higher level entities, as typical sampling frames been been restricted and limit the power of investigations to find significant cross-level interactions.

Our results clearly highlight that the power to detect cross-level interactions is severely limited in many circumstances. One response to this situation could be to "open the alpha window" and adopt more lenient levels such as $\alpha = .10$ or .15 for such tests (see Cascio & Zedeck, 1983). But we do not want to promote the adoption of new universal alpha levels that may become new standards no less arbitrary than the typical $p < .05$. Instead, we advocate that researchers advance reasoned arguments for the a priori alpha levels that they adopt in any investigation (see Aguinis, 2004). Such arguments should weigh the relative costs and benefits of Type I versus Type II errors and the maturity of the subject domain. For example, it may be reasonable to adopt more liberal alpha levels for early investigations in a nascent topic area. As the research base matures and relationships become better understood, then perhaps more stringent alpha levels should be adopted. In other instances, for example, in safety and health domains, the costs and consequences of one type of error may be dire as compared to the alternative. In those instances, grounded researchers might adopt even more extreme alpha levels (e.g., $\alpha = .01$ or .20) to balance such concerns. In short, power is a serious concern when it comes to testing cross-level interactions. But rather than simply advocating a relaxed standard for such tests, we recommend that researchers conduct a priori power analyses and then adopt a well-reasoned alpha level for their context and domain.

In sum, our research leads to the following specific guidelines in terms of designing future multilevel studies. First, researchers should use measures with high reliability and construct validity. Second, they should not rely on simple conventions such as the 30-30 rule because there are complex and interactive relationships among the factors that affect power. Third, researchers should consider the relative impact of the various factors that affect power as revealed by our Monte Carlo simulation. They should conduct a pilot study to obtain reasonable values for each, whenever feasible, or otherwise estimate likely parameter values. Fourth, they should conduct an a priori power analysis using ML Power Tool to estimate power before data are collected. They can use the ML Power Tool to understand how power will change based on various research design and measurement choices, including choices that have different cost implications in terms of time and financial resources as well as practical constraints involved in the data collection effort. Following these guidelines is likely to lead to improved accuracy concerning the presence or absence of cross-level interaction effects.

Limitations, Boundary Conditions, and Extensions

All Monte Carlo simulations are bound by the parameters selected to be manipulated and their values included in the design. Although our simulation was quite large in scope, any such effort can always include additional parameters and values. Our choices for parameters were guided by our theory-based expectations regarding factors known or hypothesized to affect the power to detect cross-level interaction effects. Moreover, our choices for parameter values were guided by ones as reported in the applied psychology literature. In spite of these considerations, we had to make some decisions about what to include and what not to include in our simulation. For example, factors known to affect the power of continuous moderators are generally similar to those known to affect the power of categorical moderators in the context of regression models (Aguinis, 2004). So, we anticipate that our results regarding the relative impact of factors that affect the power to detect continuous moderators will generalize to the case of categorical moderators in the context of multilevel models. Nevertheless, there are additional issues that are raised when the moderator is categorical, such as a resulting nonnormal distribution.

As a second example of Monte Carlo design features, our simulation held Level 1 sizes constant within each design cell and assumed homogeneity of within-group variances. Again, we do not anticipate that deviations from this scenario will affect our results in a substantive manner. Similarly, Maas and Hox (2005) found little influence of heterogeneous Level 1 sample sizes on power estimates, beyond what the mean Level 1 value suggested. Nevertheless, now that our simulations have led to knowledge regarding the relative impact of Level 1 and Level 2 sample sizes on the power to estimate cross-level interaction effects, future research could examine the extent to which severe heterogeneity of lower level sample sizes and variances across upper level units affect the power to detect existing cross-level interactions. Our simulations also adopted typical hierarchical modeling assumptions, such as normal variable distributions and uncorrelated errors within and across levels. Naturally, to the extent that such assumptions do not hold, they will likely negatively impact the power to detect true effects, the degree to which remains an issue for future investigations. Furthermore, our work—and any application of power analyses—assumes that the targeted effect size estimate is accurate. To the extent that previous investigations have yielded effect size estimates that may have been biased by threats to internal validity, endogeneity influences, or other sources of contamination, the entire exercise is compromised.

We should also highlight that we employed within-group centering and reintroduced the between-group variance in the Level 1 predictor as a Level 2 predictor. This is a commonly recommended procedure for testing cross-level interactions (cf. Enders & To-

fighi, 2007; Hofmann & Gavin, 1998) and permits one to partition the observed lower level predictor variance (and potential interactions) into that which resides within versus between Level 1 units. However, the most appropriate centering techniques for any application are guided by the researchers' underlying assumptions and theory concerning the relationships being tested (Hox, 2010). These become even more complex in the repeated measures versions of multilevel designs (Biesanz et al., 2004). Our current findings are limited to nested multilevel designs using within-group centering approach, but research is warranted to test its application to other centering approaches and designs. In a related vein, we advocate standardizing variables on the basis of the total sample for lower level variables and at the upper level for Level 2 variables. This practice is widely recommended by others who have advanced other power estimation approaches (e.g., Raudenbush & Liu, 2000), but there are costs and benefits of doing so. On one hand, Hox (2010) argued,

> When we consider power problems in multilevel designs, it is useful to translate the model under consideration into a standardized model. In a standardized model, the logic of power analysis, including decisions about effects sizes, does not depend on the arbitrary scales of the variables. In a standardized framework, it also becomes possible to establish rules of thumb about what may be considered small, medium, and large effects. (p. 243)

On the other hand, standardizing variables influences the substantive interpretation of effects in applied research and undermines direct comparisons of effects across samples or times. The larger point is that effect size estimates in multilevel investigations are dicey entities. In fact, Snijders and Bosker (1999) noted the traditional "definition of R^2 now and then leads to unpleasant surprises: it sometimes happens that adding [significant] explanatory variables increases rather than decreases some of the [error] variance components. Even negative values of R^2 are possible" (p. 99). In other words, effect sizes in multilevel research are complex phenomena that are neither totally understood nor directly comparable to their single-level analogues. The implications of centering decisions, standardization, and effect size estimates for interpreting multilevel research findings all deserve further study.

We should note that, for simplicity sake, we limited our investigations to models with one lower level predictor and one upper level predictor/moderator. However, Raudenbush (1997) and others have demonstrated that the power to detect significant lower or cross-level direct effects can be enhanced by including significant covariates from either level. The logic is that covariates help to reduce estimates of standard errors, thereby increasing power. We would surmise that including covariates in one's analysis would also enhance the power to detect cross-level interactions. However, exactly how this operates is likely to be quite complex and to be a function of the covariances among variables within and across levels, centering decisions, as well as error distributions and covariances across levels. This represents the natural next frontier for work along these lines. Another design issue that should be considered would be the role of higher level nesting. Our work has been limited to a two-level design, such as individuals in teams. However, if the teams were sampled from multiple organizations, as recommended above, there would be another level of nesting and interdependence. Given that a failure to consider salient nesting arrangements leads to underestimates of standard errors, the power values from our two-level approximation are likely to overestimate power if there are significant higher level nesting effects that are not taken into account. The extent to which higher level nesting compromises the accuracy of two-level power estimates, however, remains a topic for future investigations.

Beyond measurement, research design, and sampling issues, as well as covariate inclusion, future progress in our knowledge regarding the power to estimate cross-level interactions accurately may be garnered through advancements in analytic techniques. For example, popular hierarchical modeling software packages (e.g., HLM; Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2004) have recently incorporated robust standard error estimates that adjust for minor violations of normality assumptions. Unfortunately, Maas and Hox (2004) suggested that robust errors should only be used with relatively large numbers of Level 2 units (i.e., $\geq$ 100), which limits their use in instances where power needs are acute. Further, advancements in the domain of multilevel structural equation modeling (ML-SEM) have been prominent as of late, and they incorporate the ability to help adjust for measurement unreliability (cf., Muthén & Muthén, 2007). We anticipate that developments in that area will heighten our ability to detect cross-level interactions in future research, although ML-SEM analyses are also very sensitive to sample size issues—the very crux of power analyses.

Our findings also yield suggestions for the reporting of future multilevel investigations. First, we encourage the reporting of ICCs not only for lower level outcomes but also for lower level predictors. Second, many researchers report reliabilities of Level 2 variables that were calculated using lower level data. We encourage the reporting of variable reliabilities as aligned with the levels at which they are used in the substantive analyses (see Chen, Mathieu, & Bliese, 2004). Third, the variability of lower level slopes is rarely reported and is critical to the estimation of the power to detect cross-level interactions. These values are output by most multilevel software programs or are relatively easy to calculate, and we encourage future researchers to routinely report them.

Conclusion

Although multilevel investigations have been increasing in popularity at almost an exponential rate, our ability to interpret the meaning of cross-level interaction tests has been seriously limited by an inability to derive power estimates. We developed a means to a priori estimate the power to detect cross-level interactions based on an extensive simulation. Application of our method to a sample of previous multilevel studies illustrated that the power to detect significant cross-level interactions was extremely low. Use of ML Power Tool will better enable scholars to interpret their findings more accurately in terms of Type I and II errors and will also enable them to design more powerful multilevel studies that will enhance the accuracy of future inferences about cross-level interaction tests.

References

Aguinis, H. (2004). *Regression analysis for categorical moderators.* New York, NY: Guilford Press.

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. doi:10.1037/0021-9010.90.1.94

Aguinis, H., Boyd, B., Pierce, C. A., & Short, J. (2011). Walking new avenues in management research methods and theories: Bridging micro and macro domains. *Journal of Management, 37,* 395–403. doi:10.1177/0149206310382456

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. doi:10.1177/1094428108322641

Atwater, L., Wang, M., Smither, J. W., & Fleenor, J. W. (2009). Are cultural characteristics associated with the relationship between self and others' ratings of leadership? *Journal of Applied Psychology, 94,* 876–886. doi:10.1037/a0014561

Biesanz, J. C., Deeb-Sossa, N., Papadakis, A. A., Bollen, K. A., & Curran, P. J. (2004). The role of coding time in estimating and interpreting growth curve models. *Psychological Methods, 9,* 30–52. doi:10.1037/1082-989X.9.1.30

Bliese, P. D., Chan, D., & Ployhart, R. E. (2007). Multilevel methods: Future directions in measurement, longitudinal analysis, and nonnormal outcomes. *Organizational Research Methods, 10,* 551–563. doi:10.1177/1094428107301102

Bliese, P. D., & Jex, S. M. (1999). Incorporating multiple levels of analysis into occupational stress research. *Work & Stress, 13,* 1–6. doi:10.1080/026783799296147

Bohrnstedt, G. W., & Marwell, G. (1978). The reliability of products of two random variables. In K. F. Schuessler (Ed.), *Sociological methodology* (pp. 254–273). San Francisco, CA: Jossey-Bass.

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. User's manual (Version 2.1). Retrieved from http://stat.gamma.rug.nl/Pint21_UsersManual.pdf

Bowers, K. S. (1973). Situationism in psychology: An analysis and a critique. *Psychological Review, 80,* 307–336. doi:10.1037/h0035592

Buss, A. R. (1977). The trait–situation controversy and the concept of interaction. *Personality and Social Psychology Bulletin, 3,* 196–201. doi:10.1177/014616727700300207

Cao, J., & Ramsay, J. O. (2010). Linear mixed-effects modeling by parameter cascading. *Journal of the American Statistical Association, 105,* 365–374.

Cascio, W. F., & Zedeck, S. (1983). Open a new window in rational research planning: Adjust alpha to maximize statistical power. *Personnel Psychology, 83,* 517–526.

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. doi:10.1037/0021-9010.92.2.331

Chen, G., Mathieu, J. E., & Bliese, P. D. (2004). A framework for conducting multilevel construct validation. In F. J. Dansereau & F. Yammarino (Eds.), *Research in multi-level issues: The many faces of multi-level issues* (Vol. 3, pp. 273–303). Oxford, England: Elsevier Science.

Clarke, P., & Wheaton, B. (2007). Addressing data sparseness in contextual population research: Using cluster analysis to create synthetic neighborhoods. *Sociological Methods and Research, 35,* 311–351.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cooper, W. H., & Richardson, A. J. (1986). Unfair comparisons. *Journal of Applied Psychology, 71,* 179–184. doi:10.1037/0021-9010.71.2.179

Cortina, J. M., & DeShon, R. P. (1998). Determining relative importance of predictors with the observational design. *Journal of Applied Psychology, 83,* 798–804. doi:10.1037/0021-9010.83.5.798

Culpepper, S. A. (2010). Studying individual differences in predictability with gamma regression and nonlinear multilevel models. *Multivariate Behavioral Research, 45,* 153–185. doi:10.1080/00273170903504885

Culpepper, S. A., & Aguinis, H. (2011). R is for revolution: A review of a cutting-edge, free, open source statistical package. *Organizational Research Methods, 14,* 735–740. doi:10.1177/1094428109355485

Dierdorff, E. C., Rubin, R. S., & Morgeson, F. P. (2009). The milieu of managerial work: An integrative framework linking work context to role requirements. *Journal of Applied Psychology, 94,* 972–988. doi:10.1037/a0015456

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. doi:10.1037/1082-989X.12.2.121

Endler, N. S., & Magnusson, D. (1976). Toward an interactional psychology of personality. *Psychological Bulletin, 83,* 956–974. doi:10.1037/0033-2909.83.5.956

Griffin, M. A. (2007). Specifying organizational contexts: Systematic links between contexts and processes in organizational behavior. *Journal of Organizational Behavior, 28,* 859–863. doi:10.1002/job.489

Grizzle, J. W., Zablah, A. R., Brown, T. J., Mowen, J. C., & Lee, J. M. (2009). Employee customer orientation in context: How the environment moderates the influence of customer orientation on performance outcomes. *Journal of Applied Psychology, 94,* 1227–1242. doi:10.1037/a0016404

Hitt, M. A., Beamish, P. W., Jackson, S. E., & Mathieu, J. E. (2007). Building theoretical and empirical bridges across levels: Multilevel research in management. *Academy of Management Journal, 50,* 1385–1399. doi:10.5465/AMJ.2007.28166219

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.

Hofmann, D. A., Morgeson, F. P., & Gerras, S. J. (2003). Climate as a moderator of the relationship between leader–member exchange and content specific citizenship: Safety climate as an exemplar. *Journal of Applied Psychology, 88,* 170–178. doi:10.1037/0021-9010.88.1.170

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

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Klein, K. J., Cannella, A., & Tosi, H. (1999). Multilevel theory: Challenges and contributions. *Academy of Management Review, 24,* 243–248.

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

Liao, H., & Rupp, D. E. (2005). The impact of justice climate and justice orientation on work outcomes: A cross-level multifoci framework. *Journal of Applied Psychology, 90,* 242–256. doi:10.1037/0021-9010.90.2.242

Longford, N. T. (1993). *Random coefficient models.* Reading, MA: Addison-Wesley.

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. doi:10.1016/j.csda.2003.08.006

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1,* 86–92. doi:10.1027/1614-2241.1.3.85

Mathieu, J., Ahearne, M., & Taylor, S. R. (2007). A longitudinal cross-level model of leader and salesperson influences on sales force technology use and performance. *Journal of Applied Psychology, 92,* 528–537. doi:10.1037/0021-9010.92.2.528

Mathieu, J. E., & Chen, G. (2011). The etiology of the multilevel paradigm in management research. *Journal of Management, 37,* 610–641. doi:10.1177/0149206310364663

Mathieu, J. E., & Taylor, S. (2007). A framework for testing meso-

mediational relationships in organizational behavior. *Journal of Organizational Behavior, 28,* 141–172. doi:10.1002/job.436

Mischel, W. (2004). Toward an integrative science of the person. *Annual Review of Psychology, 55,* 1–22. doi:10.1146/annurev.psych.55.042902.130709

Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide* (3rd ed.). Los Angeles, CA: Author.

Neal, A., & Griffin, M. A. (2006). A study of the lagged relationsip among safety climate, safety motivation, safety behavior, and accidents at the individual and group levels. *Journal of Applied Psychology, 91,* 946–953. doi:10.1037/0021-9010.91.4.946

Pervin, L. A. (1989). Persons, situations, interactions: The history of a controversy and a discussion of theoretical models. *Academy of Management Review, 14,* 350–360.

Ployhart, R. E., & Vandenberg, R. J. (2010). Longitudinal research: The theory, design, and analysis of change. *Journal of Management, 36,* 94–120. doi:10.1177/0149206309352110

R Development Core Team. (2011). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Available at http://www.R-project.org/

Raudenbush, S. W. (1989a). "Centering" predictors in multilevel analysis: Choices and consequences. *Multilevel Modeling Newsletter, 1*(2), 10–12.

Raudenbush, S. W. (1989b). A response to Longford and Plewis. *Multilevel Modeling Newsletter, 1*(3), 8–10.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trails. *Psychological Methods, 2,* 173–185. doi:10.1037/1082-989X.2.2.173

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

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T., & du Toit, M. (2004). *HLM6: Hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5,* 199–213. doi:10.1037/1082-989X.5.2.199

Rousseau, D. M. (1985). Issues of level in organizational research: Multi-level and cross-level perspectives. *Research in Organizational Behavior, 7,* 1–37.

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. doi:10.1177/1094428107308906

Schneider, B. (1983). Interactional psychology and organizational behavior. *Research in Organizational Behavior, 5,* 1–31.

Schneider, B., Goldstein, H. W., & Smith, B. (1995). The ASA framework: An update. *Personnel Psychology, 48,* 747–773. doi:10.1111/j.1744-6570.1995.tb01780.x

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components.* New York, NY: Wiley.

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, pp. 1570–1573). Chichester, England: Wiley.

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18,* 237–259. doi:10.2307/1165134

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London, England: Sage.

Spybrook, J., Raudenbush, S. W., Congdon, R., & Martinez, A. (2009). Optimal design for longitudinal and multilevel research: Documentation for the "Optimal Design" Software (Version 2.0). Retrieved from http://www.wtgrantfoundation.org/resources/overview/research_tools

Wallace, J. C., Edwards, B. D., Arnold, T., Frazier, M. L., & Finch, D. M. (2009). Work stressors, role-based performance, and the moderating influence of organizational support. *Journal of Applied Psychology, 94,* 254–262. doi:10.1037/a0013090

Yu, K. Y. T. (2009). Affective influences in person–environment fit theory: Exploring the role of affect as both cause and outcome of P-E fit. *Journal of Applied Psychology, 94,* 1210–1226. doi:10.1037/a0016403

Zhang, D., & Willson, V. L. (2006). Comparing empirical power of multilevel structural equation models and hierarchical linear models: Understanding cross-level interactions. *Structural Equation Modeling, 13,* 615–630. doi:10.1207/s15328007sem1304_6

(*Appendix follows*)

Appendix

Multilevel (ML) Power Tool

```
l2n = 62 #Level-2 sample size
l1n = 7 #Average Level-1 sample size
iccx = .12 #ICC1 for X
g00 = −0.068364 #Intercept for B0j equation (Level-1 intercept)
g01 = 0.345048 #Direct cross-level effect of average Xj on Y
g02 = 0.022851 #Direct cross-level effect of W on Y
g03 = 0.184721 #Between-group interaction effect between W and Xj on Y
g10 = 0.451612 #Intercept for B1j equation (Level-1 effect of X on Y)
g11 = 0.148179 #Cross-level interaction effect
vu0j = 0.00320 #Variance component for intercept
vu1j = 0.08954 #SD of Level-1 slopes
vresid = 0.76877 #Variance component for residual, within variance
alpha = .05 #Rejection level
REPS = 1000 #Number of Monte Carlo replications, 1,000 recommended

hlmmmr <-function(iccx,l2n,l1n,g00,g01,g02,g03,g10,g11, vu0j,vu1j,alpha){
require(lme4)
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. Users supply underlined values. This example uses input data from Chen et al. (2007). Syntax must be run in R (R Development Core Team, 2011) including the linear mixed-effects models using S4 classes (lme4) module. This syntax and further information is available at http://mypage.iu.edu/~haguinis/crosslevel.html

Received May 14, 2010
Revision received February 20, 2012
Accepted March 1, 2012 ■

See page 981 for a correction to this article.

Correction to Mathieu, Aguinis, Culpepper, and Chen (2012)

The article "Understanding and Estimating the Power to Detect Cross-Level Interaction Effects in Multilevel Modeling," by John E. Mathieu, Herman Aguinis, Steven A. Culpepper, and Gilad Chen (*Journal of Applied Psychology*, Advance online publication. May 14, 2012. doi:10.1037/a0028380), contained production-related errors in a number of the statistical symbols presented in Table 1, the Power in Multilevel Designs section, the Simulation Study section, and the Appendix. All versions of this article have been corrected.

DOI: 10.1037/a0029358