Multilevel Modeling: Research-Based Lessons for Substantive Researchers

Vicente González-Romá and Ana Hernández

Research Institute of Personnel Psychology, Organizational Development, and Quality of Working Life (Idocal), Faculty of Psychology, University of Valencia, 46010 Valencia, Spain; email: vicente.glez-roma@uv.es, ana.hernandez@uv.es

Annu. Rev. Organ. Psychol. Organ. Behav. 2017. 4:183–210

First published online as a Review in Advance on January 11, 2017

The *Annual Review of Organizational Psychology and Organizational Behavior* is online at orgpsych.annualreviews.org

https://doi.org/10.1146/annurev-orgpsych-041015-062407

Keywords

multilevel modeling, multilevel structural equation modeling, multilevel mediation, multilevel moderation, contextual effects

Abstract

Organizations are multilevel systems. Most organizational phenomena are multilevel in nature, and their understanding involves variables (e.g., antecedents and consequences) that reside at different levels. The investigation of these phenomena requires appropriate analytical methods: multilevel modeling. These techniques are becoming increasingly popular among organizational psychology and organizational behavior (OPOB) researchers. In this article we review the literature that has evaluated the performance of multilevel modeling techniques to test multilevel direct and indirect effects and cross-level interactions. We also provide guidelines for OPOB researchers about the appropriate use of these techniques, and we suggest ways these techniques can contribute to theoretical advancement and research development in OPOB.

INTRODUCTION

Organizations are multilevel systems in which organizational entities (e.g., employees, teams, departments, organizations) reside in nested arrangements (e.g., employees are nested in teams, teams in departments, and departments in organizations) (Kozlowski & Klein 2000, Mathieu & Chen 2011). Each position in these nested arrangements denotes a specific level.

In the twentieth century, organizational researchers tended to focus on single-level analyses, disregarding the relationships among characteristics of organizational entities that reside at different levels (Hitt et al. 2007, Kozlowski & Klein 2000). This single-level perspective is limited because it cannot explain the complexities of most organizational phenomena, where the antecedents, mediators, moderators, and outcomes involved reside at different levels (Kozlowski & Klein 2000).

Fortunately, in the last 20 years of the twentieth century, the seeds for change were sown by several scholars who introduced the multilevel paradigm in organizational psychology and organizational behavior (OPOB) (e.g., House et al. 1995, Klein et al. 1994, Kozlowski & Klein 2000, Rousseau 1985). A key idea in this paradigm is that the characteristics of a given entity (e.g., a work team) are related to the characteristics of other entities that reside at different levels. For instance, the culture of an organization promotes certain human resource management practices that, once implemented in a given team by a specific leader, induce a certain team climate, which in turn produces certain levels of job satisfaction in the employees within this team. To estimate relationships that span across levels, new statistical methods were needed. Practically in parallel to the emergence of the multilevel paradigm in OPOB, several statisticians began to develop new multilevel modeling methods (also known as hierarchical linear models, mixed-effects models, random effects models, and random coefficient models) (Bryk & Raudenbush 1992, Burstein et al. 1978, de Leeuw & Kreft 1986, Goldstein 1986, Muthén 1989). This methodological development was accompanied by software that facilitated the implementation of multilevel modeling techniques in applied research. As a result of the confluence in time between these two streams of academic work, the number of multilevel studies in the field began to increase steadily from the turn of the century on. For instance, in the 1990s, the number of documents (i.e., articles, reviews, book chapters, or discussions) per year with the subject "multilevel" or "multi-level" indexed in the Web of Science categories of "Psychology, Applied", "Management," and "Business" varied between 1 (in 1990) and 22 (1995). In 2010, this figure was 188, and it was 342 in 2015. This upward trend suggests that researchers in our field have acknowledged that if most organizational phenomena are multilevel in nature, appropriate tools have to be used to investigate them. Consequently, multilevel modeling techniques are becoming increasingly popular among OPOB researchers. Thus, it makes sense to take stock and see what the research on these techniques has found out in recent decades about their performance under different conditions. This information will help OPOB researchers design and plan their multilevel investigations, taking into account the scientific evidence accumulated so far. Therefore, this article aims to (*a*) review the literature that has evaluated the functioning of multilevel modeling techniques to test direct and indirect cross-level effects and cross-level interactions, (*b*) provide guidelines for the appropriate use of these techniques for OPOB researchers who need to use them, and (*c*) suggest ways these techniques can contribute to theoretical advancement and research development in OPOB.

This article is structured as follows. First, we briefly explain why multilevel modeling techniques are needed, and we present the consequences of not using them when they should be used. Second, we focus on the conventional multilevel modeling (CMLM) techniques that evolved from multiple regression and their modifications, and we review the literature that has examined their performance. Third, we focus on multilevel structural equation modeling (MSEM). We briefly present the limitations of the CMLM techniques that MSEM overcomes, and we review the literature that has investigated its performance. Finally, we provide a set of practical guidelines for OPOB researchers, and we suggest ways that multilevel modeling techniques can contribute to knowledge advancement in OPOB. Due to space limitations and to be consistent with most of the studies conducted, we mainly focus on designs with two levels in which organizational entities can be neatly nested. The application of multilevel modeling techniques to longitudinal data (e.g., Heck et al. 2013) is beyond the scope of this review.

THE CONSEQUENCES OF DISREGARDING THE NESTED STRUCTURE OF ORGANIZATIONAL DATA

Ignoring the nested structure of data and analyzing data at the lower level by means of ordinary least squares (OLS) regression can have undesirable consequences (Heck & Thomas 2015). An important problem with this practice is related to the OLS regression assumption of independence of observations, which is violated in the case of nested data (Preacher et al. 2011). Employees who are members of the same organizational subunit tend to have similar perceptions, affects, attitudes and behaviors (González-Romá & Hernández 2014). Therefore, their responses to instruments designed to measure the variables of interest will also tend to be similar. Thus, nested data generally show some degree of nonindependence. This degree of nonindependence is indicated by the intraclass correlation coefficient [ICC(1)]. Analyzing nonindependent nested data by means of OLS regression at the lower level leads to Type I and Type II errors (Bliese & Hanges 2004). These authors have shown that when a researcher investigates the relationship between a higher-level variable (e.g., team age diversity) and a lower-level variable (e.g., individual job tension) under these conditions, the standard error (SE) of the involved parameter estimate is underestimated. Thus, the corresponding *t* ratio (i.e., parameter estimate/SE) used to test for statistical significance is inflated, leading to an increase in the probability of rejecting the null hypothesis when it is true (i.e., an increase in Type I error). Bliese & Hanges (2004) also showed that when a researcher investigates the relationship between two lower-level variables (e.g., role conflict and job tension) in nested data that show nonindependence through OLS regression, the SE of the corresponding parameter estimate is overestimated, leading to an increase in Type II error and a loss of statistical power. In a simulation study, Bliese & Hanges (2004) provided empirical evidence supporting this latter consequence. Their results showed that the degree of nonindependence in the outcome variable played a critical role: Power loss was generally greater for OLS regression when ICC(1) values in the outcome were higher rather than lower. Thus, even when the interest is in relationships at the lower level but nested data are being analyzed, not using adequate multilevel techniques can lead to loss of statistical power. Similar consequences have been observed in three-level designs where one of the higher levels was ignored (Moerbeek 2004).

The use of single-level structural equation modeling (SEM) to analyze nested data has similar problems. Finch & French (2011) showed that this practice yields underestimated SEs and an inflated Type I error, problems that become more severe as ICC(1) values increase. Therefore, even when nonindependence in nested data seems to be small [as indicated by ICC(1) values slightly greater than 0.05], researchers should analyze these data by using multilevel modeling techniques (Julian 2001). However, simulation studies suggest that when ICC(1) $\leq$ 0.05 the consequences of ignoring the nested structure of data are negligible (Finch & French 2011, Julian 2001).

CONVENTIONAL MULTILEVEL MODELING

Brief Introduction to Conventional Multilevel Modeling

When relating two individual or level-1 (L1) variables (X and Y), CMLM assumes that the regression of Y on X for a given set of sampled individuals (i) can vary depending on the level-2 (L2) work unit (j) (i.e., team, department, organization) to which they belong. There may be differences across units in the regression intercepts and in the regression slopes that capture the strength and direction of the L1 relationship between X_{ij} and Y_{ij}. The L1 equation to predict Y_{ij} from X_{ij} is

$$L1 : Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}, \tag{1}$$

where β_{0j} and β_{1j} are the random intercepts and slopes that are allowed to vary across groups, respectively, and e_{ij} is the L1 residual term.

When researchers expect the regression intercepts to differ across groups, they typically investigate whether there are L2 variables that contribute to explaining these differences, focusing on the so-called cross-level direct effects. When researchers expect the regression slopes to differ across groups, they pay attention to potential L2 variables that might explain the differences in strength and/or direction of the L1 relationship, focusing on the so-called cross-level moderations or cross-level interactions. Thus, the intercepts and slopes become the outcomes of the L2 equations that are regressed on the L2 predictor of interest (W_j):

$$L2 : \beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \tag{2}$$

$$L2 : \beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j}, \tag{3}$$

where the L2 intercepts, γ_{00} and γ_{10}, and slopes, γ_{01} and γ_{11}, are fixed coefficients, and u_{0j} and u_{1j} are the random intercept and random slope residuals, respectively. Plugging the L2 equations into L1, we have the following combined equation:

Erratum

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} W_j + \gamma_{11} X_{ij} W_j + u_{0j} + u_{1j} X_{ij} + e_{ij}, \tag{4}$$

with γ_{10} and γ_{01} estimating the main effects of the L1 and L2 predictors, respectively, and γ_{11} estimating the interaction between them.

Focusing on cross-level direct effects, Naumann & Bennett (2000) showed that procedural justice climate (an L2 predictor) was positively related to helping behaviors (an L1 outcome), once the effects of individual procedural justice perceptions were controlled for. This relationship is depicted in **Figure 1*a***, with the blue arrow capturing the cross-level direct effect (γ_{01} in Equations 2 and 4). Had the relationship between the L1 variables varied across groups, an L2 moderator could have been introduced to explain these differences. **Figure 1*b*** shows a cross-level moderation, on the basis of Lee & Dalal's (2016) study. They showed that organizational safety climate strength moderates the L1 relationship: Conscientiousness → Safety behavior. **Figure 1*b*** represents this effect (γ_{11} in Equations 3 and 4) by means of the blue arrow that impacts the random coefficient b_j. If the interaction effect is statistically significant, then the conditional effects should be estimated, plotted, and tested using the equations derived by Bauer & Curran (2005) (see also Preacher et al. 2006; for a list of related resources about this and other multilevel topics, follow the **Supplemental Material link** in the online version of this article or at **http://www.annualreviews.org/**). For a comprehensive review on best practices for estimating and testing cross-level moderation effects, see Aguinis et al. (2013).

Supplemental Material

Nevertheless, apart from moderation, OPOB researchers are typically interested in understanding the mediating mechanisms that explain the relationships among the focal variables. When the mediator is located at different levels from the predictor and/or the outcome, we have cross-level

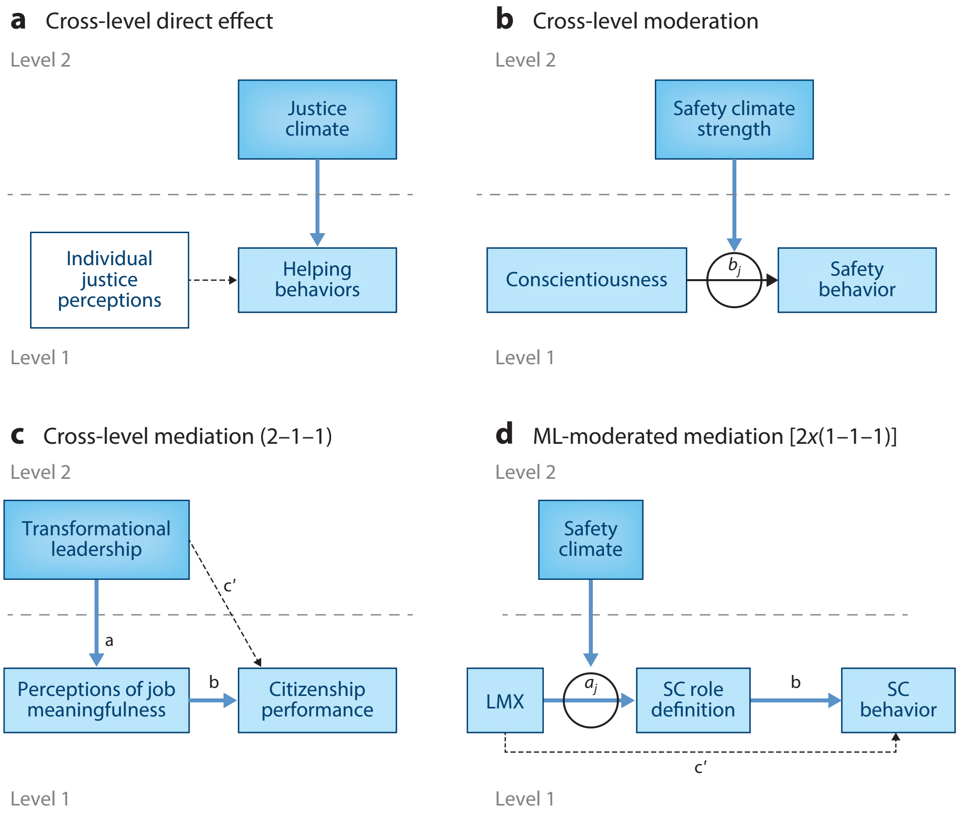

Figure 1

Multilevel effects. (*a*) Cross-level direct effect. (*b*) Cross-level moderation. (*c*) Cross-level mediation (2–1–1). (*d*) ML-moderated mediation [2*x*(1-1-1)]. Arrows without circles represent fixed effects. Arrows with circles represent random effects. Depiction of residuals has been omitted for simplicity. Diagrams are conceptual and do not exactly correspond to the statistical model to be tested. Abbreviations: LMX, leader-member exchange; ML, multilevel; SC, safety citizenship.

mediation. We may have a 2–1–1 or a 2–2–1 mediation (the numbers indicate the level of the predictor, mediator and outcome variables, respectively; Krull & MacKinnon 2001). In **Figure 1**, panel *c*, we have depicted a 2–1–1 model based on Purvanova et al. (2006): “Transformational leadership → Employees’ job perceptions (e.g., meaningfulness) → Employees’ citizenship performance.” The mediated or indirect effect is obtained from the product of the coefficients involved in the mediation (see MacKinnon et al. 2007) (*ab* in panel *c* of **Figure 1**). The first coefficient (*a*) is obtained by regressing the L1 mediator (M_{ij}) on the L2 predictor (X_j) in the corresponding intercept-L2 equation (an equation similar to Equation 2). The second coefficient (*b*) is obtained by regressing the L1 outcome (Y_{ij}) on the L1 mediator, controlling for the effect of the L2 predictor (X_j), introduced in the corresponding intercept-L2 equation. As in single-level mediation, the statistical significance of the *ab* product must be tested using appropriate methods (see Pituch & Stapleton 2008 for a review).

Finally, if one of the L1 paths involved in a mediated relationship differs across groups, researchers can look for potential moderators of these differences. In this case, a multilevel moderated mediation is tested. This is the case of our final example. Based on Hofmann et al. (2003),

Figure 1, panel *d* shows a 1–1–1 multilevel mediation, "Leader–Member Exchange (LMX) → Safety citizenship role definitions → Safety citizenship behavior," where the first path, the random coefficient a_j, is moderated by team safety climate. In this case, the magnitude of the indirect effect (for the example, $\overline{a_j}b$) is expected to depend on the L2 moderator. Kenny et al. (2003) and Bauer et al. (2006) show a general strategy to test for moderated multilevel mediation using a 1–1–1 model with an L2 variable moderating all individual relationships (including the effect of X_{ij} on Y_{ij} after controlling for M_{ij}). If any of the interaction terms involved in the mediation path are statistically significant, the conditional indirect effects at different levels of the moderator can be tested by means of simple indirect effects (see Bauer et al. 2006).

Of course, the examples in **Figure 1** represent only some of the possible multilevel models that may be of interest in OPOB. For more in-depth tutorials on the logic and rationale of different models that can be tested using CMLM, see Hofmann (1997) and Krull & MacKinnon (1999, 2001).

Theoretical versus Statistical Interpretation of Relevant Effects in Conventional Multilevel Modeling

Although multilevel theories may consider cross-level upward influences, in CMLM the outcome variable Y_{ij} is always defined at the lower level. Thus, a cross-level direct effect theoretically refers to how variations in an L2 predictor relate to variations in an L1 outcome (e.g., **Figure 1*a***). This type of hypothesis, which is common in the OPOB literature, warrants further explanation.

The combined Equation 4 shows that the L2 predictor W_j has an effect on the L1 outcome Y_{ij}. And it does. In our example, part of the variability of Y_{ij} can be attributed to justice climate. However, this effect is assumed to be constant for all individuals belonging to the same work unit, and it cannot influence individual differences within a group. Cross-level effects are not within-group effects. They are between-group effects. As LoPilato & Vandenberg (2015) point out, "the theoretical cross-level direct effect is different from the statistical direct effect" (p. 301). The theoretical interpretation will be accurate only to the extent that the within-unit variability in the outcome is null or very low; therefore, the group intercept (i.e., the group mean) is a good proxy for the individual score. Thus, researchers should be careful when writing their conclusions about cross-level direct effects. For a more in-depth discussion of these issues and recommended steps for testing cross-level direct effects, see LoPilato & Vandenberg (2015).

Regarding cross-level moderators, they are typically defined as higher-level variables that change the nature and/or strength of the within-group relationship between two L1 variables. In the cross-level moderation in our example (**Figure 1*b***), the hypothesis is that the relationship between employees' conscientiousness and safety behaviors varies as a function of organizational safety climate strength. However, the definition does not fit the effect tested by the equations presented. In nested data, individual scores tend to show some degree of dependency due to group membership; therefore, L1 variables typically share both within-group and between-group variance. Consequently the interaction term γ_{11} conflates two types of effects: the theoretical cross-level interaction of interest here, $X_{ij}W_j$, and the upper-level (i.e., between-group) interaction, X_jW_j , where X_j represents the average unit score on X_{ij} (Aguinis et al. 2013, Enders & Tofighi 2007, Hofmann & Gavin 1998). This is the problem of conflated variance, which is not exclusive to cross-level moderation.

For 2–2–1 mediation, conflated variance is not an issue because X and M are L2 variables. However, the same considerations presented above for cross-level direct effects are relevant for the 2–2–1 cross-level indirect effect *ab*. For 2–1–1 mediation models, because the coefficient that relates the mediator to the outcome is a mixture of between and within variance, the indirect effect

($a\overline{b_j}$) also conflates both types of effects (Zhang et al. 2009). Finally, regarding 1–1–1 models, the two coefficients involved in the indirect effect are typically a mixture of between- and within-group variance. When these two coefficients vary across groups, their covariance should be incorporated into the estimate of the indirect effect (see Kenny et al. 2003 and Bauer et al. 2006 for two proposals about how to deal with this issue in CMLM). For a more-in-depth discussion about how variance is partitioned into different sources, see Aguinis et al. (2013).

In the next section, we focus on possible strategies to solve the interpretation problems of multilevel regression coefficients, especially when L1 predictors conflate both between- and within-group effects.

Rescaling Predictors and Disentangling Within and Between Effects

Multilevel Equations 1 to 4 have been presented in raw-score form. However, the interpretation of regression coefficients by using raw scores is problematic in OPOB research because a value of zero usually does not have a meaningful interpretation. Thus, it is necessary to rescale the L1 and L2 predictors. In general, the recommendation is to rescale L1 predictors by using group mean centering to obtain an accurate estimate of within-group slopes (e.g., Dalal & Zickar 2012, Enders & Tofighi 2007, Hofmann & Gavin 1998, Bryk & Raudenbush 1992, Zhang et al. 2009). However, as we describe below in this section, this recommendation must be qualified. When L2 predictors do not have a meaningful zero, they should also be rescaled. In this case, grand-mean centering (GMC) is the choice, unless the predictor is a dummy variable or there are reasons to use a specific arbitrary value (Aguinis et al. 2013, Enders & Tofighi 2007). Following Enders & Tofighi (2007), we use the expression centering within cluster (CWC) for group-mean centering, to avoid confusion with GMC. These two centering options do not change the CMLM statistical model.

If L1 predictors are not rescaled, or they are rescaled by GMC, scores are an indistinguishable mixture of within-group and between-group variance (Bryk & Raudenbush 1992, Hofmann & Gavin 1998, Paccagnella 2006, Zhang et al. 2009). By using the CWC option, the between variation is removed, and the problem of conflated variance is avoided: The L1 random slopes are within-group coefficients that accurately represent the within-group relationship between X_{ij} and Y_{ij} (Aguinis et al. 2013, Dalal & Zickar 2012, Hofmann & Gavin 1998). However, variables that exclusively explain within-group variance are rare in nested data. To take into account the possible between effect, the standard recommendation is to bring back the between-group variance. This is achieved by modifying the CMLM statistical model and introducing the group mean (X_j) into the model as an L2 predictor (e.g., Hofmann & Gavin 1998, Paccagnella 2006, Zhang et al. 2009). This L2 variable may also be rescaled by using GMC (Enders & Tofighi 2007, Mathieu et al. 2012). The strategy of using CWC and reintroducing the means of the involved predictors at L2 [denoted as CWC(M) (Zhang et al. 2009)] results in the unconflated multilevel model [UMM (Preacher et al. 2010, 2011)]. This model allows researchers to test whether the association between X and Y is different at both levels of the hierarchy and whether, consequently, a contextual effect exists (i.e., the group means provide additional explanatory power) (Enders & Tofighi 2007). This is also possible when using GMC for L1 and introducing the group mean M_j at L2 to control for the between effect [the strategy referred to as GMC(M)]. Adding the group means as a predictor serves to partial out the L2 influence of the predictor, resulting in an unbiased estimate of the L1 regression slope (Kreft et al. 1995, Enders & Tofighi 2007).

For cross-level interactions (e.g., **Figure 1*b***), if X_{ij} (conscientiousness) had both a within and a between effect on Y_{ij} (safety behavior), and these effects were moderated by W_j (climate strength), GMC of X_{ij} would result in an interaction term that conflates both types of effects, whereas CWC would ignore the between-group effect. The CWC(M) and GMC(M) options allow researchers to differentiate the cross-level interaction $X_{ij}W_j$ (whether the within-group relationship between conscientiousness and helping behavior depends on climate strength) from the between interaction X_jW_j (whether the between-group relationship between group conscientiousness and group helping behavior depends on climate strength) [see Enders & Tofighi 2007 for a detailed explanation of the relationship between CWC(M) and GMC(M) and how to test for contextual effects in both cases].

When focusing on multilevel mediation, we have different scenarios. For 1–1–1 mediation, we can, by using CWC(M) or GMC(M), differentiate how much of the indirect effect is between and how much is within, as well as test for contextual effects. For 2–2–1 mediation, conflated variance is not an issue because all the involved relationships refer to between effects (Zhang et al. 2009).

Finally, for 2–1–1 models, there is some controversy about the statistical versus theoretical interpretation of the indirect effect. Some researchers argue that "any mediation of the effect of a Level-2 X must also occur at a between-group level, regardless of the level at which M and Y are assessed, because the only kind of effect that X can exert (whether direct or indirect) must be at the between-group level" (Preacher et al. 2010, p. 210; see, also, Zhang et al. 2009). Hence, for a 2–1–1 model like the one **Figure 1*c*** depicts, any direct or indirect effect of X exists only between groups. If raw scores or GMC were used for the L1 mediator, the b coefficient involved in the indirect effect would conflate both between- and within-group effects. By using the UMM [i.e., the CWC(M)], the indirect between-group effect of X on Y through M can be estimated unequivocally. Even if the within effect $[(M_{ij} - M_j) \rightarrow Y_{ij}]$ is also estimated, it is irrelevant for obtaining the indirect effect (Zhang et al. 2009). Using simulated data, Zhang et al. (2009) showed that CWC(M) avoided the problem of conflated variance in 2–1–1 models, whereas GMC did not.

In the context of cluster randomized trials, some researchers (Pituch & Stapleton 2012, Tofighi & Thoemmes 2014) argue that if the mediator is rescaled by GMC(M), then the indirect effect in 2–1–1 models can be decomposed into a cross-level indirect effect and a between-group indirect effect. The first path involved in the mediation (a) is the same for both types of indirect effects because the effect of X_j on M_{ij} is constant for all individuals belonging to a group. For the second path, two effects are differentiated: the within effect of M_{ij} on Y_{ij} (b_w) and the between effect of M_j on Y_{ij} intercepts (b_b). The product ab_w is an estimate of the cross-level indirect effect (provided that, as for cross-level direct effects, the assumption of a constant effect of X_j on M_{ij} is reasonable). The product ab_b is an estimate of the between-group indirect effect. The total indirect effect is the sum of these two effects ($ab_w + ab_b$). If CWC(M) is used, then Pituch & Stapleton (2012) and Tofighi & Thoemmes (2014) agree with Preacher et al. (2010) that the only indirect effect that can be estimated is the between effect. Using simulated data, Pituch & Stapleton (2012) compared the performance of GMC, GMC(M) and CWC(M). When no contextual effects were present for the mediator, the statistical power was four times lower for CWC(M) (0.18 on average) than for GMC and GMC(M). When contextual effects were present for the mediator, the statistical power was always greater for the cross-level indirect effect, even when the between-group indirect effect was larger (probably because more information is typically available at the individual level than at the group level). Thus, Pituch & Stapleton (2012) conclude that "there appears to be an important practical benefit of estimating the two separate indirect effects " (p. 659), which are available when using GMC(M) for the L1 mediator.

However, in spite of the practical issues, centering decisions should be based on the theoretical processes researchers want to test (Aguinis et al. 2013, Hofmann & Gavin 1998, Pituch & Stapleton 2012). Moreover, the interpretation of results should always reflect the centering method used. When focusing on frog-pond or social comparison processes, CWC(M) is a better choice. This is what Hofmann et al. (2003) did in the LMX example in **Figure 1*d***. In other cases, "it may be more appropriate to use grand-mean centering with across-group variance controlled because a theory may address raw differences between L1 entities, not differences relative to a group average" (Aguinis et al. 2013, p. 23). This may be the case of Lee & Dalal's (2016) example on conscientiousness (**Figure 1*b***), which compared the results obtained using different centering options (see other interesting examples in Lüdtke et al. 2009). In fact, using and comparing different centering options may increase our understanding of some multilevel relationships. For a more in-depth examination of different interpretations of parameter estimates depending on the type of centering, see Hofmann & Gavin (1998), Enders &Tofighi (2007), Lüdtke et al. (2009), and Enders (2013).

Estimating and Testing Multilevel Effects

A relevant issue is to determine under what conditions researchers can obtain adequate estimates of multilevel fixed effects. However, residual random components also deserve attention.

Residual variances. So far, we have paid attention to the main effects of interest in testing OPOB hypotheses. Nevertheless, the variance estimates of residual random components provide useful information when analyzing multilevel data.

First, having enough variability across intercepts and/or slopes is a precondition for including L2 predictors in CMLM (e.g., Gavin & Hofmann 2002). This variability is typically assessed by testing whether the intercept and slope variances differ significantly from zero, or by quantifying the proportion of criterion variance attributed to group membership (i.e., intercept differences) by means of ICC(1). However, both practices are problematic.

On the one hand, the SEs of the L2 residual variances are inaccurate (e.g., Maas & Hox 2004a,b; Mok 1995; Van der Leeden et al. 1997), and null-hypothesis tests for L2 residual variances have low statistical power, regardless of the specific test used (Berkhof & Snijders 2001, LaHuis & Ferguson 2009, Scheipl et al. 2008; for a brief summary of the simulation studies reviewed, follow the **Supplemental Material link** in the online version of this article or at **http://www.annualreviews.org/**). In fact, several researchers have warned that the lack of significant variance in L2 components should not prevent researchers from testing their cross-level hypotheses (e.g., Aguinis et al. 2013, LaHuis & Ferguson 2009, Snijders & Bosker 1999).

Supplemental Material

On the other hand, ICC(1) ignores the proportion of criterion variance attributed to slope differences. To solve this problem, Aguinis & Culpepper (2015) proposed ICC(β). Using empirical and simulated data, they showed that there are cases where ICC(1) is zero, suggesting that there is no need for multilevel modeling (MLM), when, in fact, there is considerable variability attributed to slope differences. Using simulated data, they also showed that ICC(β) performs better than several statistical tests in detecting slope variability. Thus, the combined use of ICC(1) and ICC(β) can provide useful information about the level and type of effect researchers should focus on. Values of ~0.05 in ICC(1) (Heck et al. 2013, LeBreton & Senter 2008) and ICC(β) (Aguinis & Culpepper 2015) may be large enough to have implications for multilevel theory and research.

Second, residual variances are useful to quantify the importance of multilevel effects by estimating effect sizes as percentages of explained variance (e.g., LaHuis et al. 2014, Selya et al. 2012). This is important if we consider that the power to detect multilevel effects is typically low, as we discuss below.

Multilevel fixed effects. A relevant issue is to determine what combinations of L2/L1 sample sizes prevent estimate bias and type I errors and foster estimates' precision and statistical power.

Simulation studies addressing these issues have consistently shown that estimates of cross-level direct effects and interactions are reasonably unbiased and precise, and they have interval confidence coverages close to the nominal value, even when the assumption of normally distributed residuals does not hold, provided that the number of groups is not too small (e.g., Bell et al. 2014; Maas & Hox 2000, 2004b, 2005; Mok 1995; Van der Leeden et al. 1997). Specifically, some of these studies report unbiased parameter estimates with as few as 20 groups of approximately 10 (Bell et al. 2014, Van der Leeden et al. 1997) or 10 groups of 5 (Maas & Hox 2005). However, to obtain reasonable SEs, a minimum of 30 groups is necessary (Maas & Hox 2004a, 2005).

Regarding multilevel indirect effects, simulation studies using raw scores or GMC in CMLM have concluded that the estimates of the indirect effects and their SEs are generally accurate (Bauer et al. 2006, Krull & MacKinnon 1999, 2001). For example, for a 2–1–1 model, Krull & MacKinnon (1999) concluded that the indirect effect *ab* had no substantial bias, even for 10 groups of 5–10 individuals. In addition, the relative bias in the SEs was also small when estimated by the first-order Taylor approximation (Sobel 1982). Regarding type I errors, Pituch & Stapleton (2008) tested the indirect effect of a 2–1–1 model by comparing several methods that take into account that the indirect effect *ab* is not normally distributed. They concluded that the bias-corrected parametric bootstrap and the empirical-M test were the best options. However, in these studies, between and within effects were conflated. Zhang et al. (2009), who argue that the indirect effect in a 2–1–1 model must be restricted to the between effect, compared GMC and the UMM in a simulation study. They concluded that point estimates of the true indirect effect (i.e., the between effect) were biased under GMC, and the type I error rates were too large. The UMM performed well for all sample size combinations (from 120/5 to 20/30). Setting aside the debate about whether a 2–1–1 indirect effect is exclusively a between effect, research shows that multilevel effects are estimated accurately with relatively small samples. However, when we focus on statistical power, the results are not quite as encouraging.

Initial studies focusing on the power of CMLM for detecting cross-level direct effects and interactions showed that the number of groups was more important than the number of individuals per group (Bassiri 1988; Kim 1990; R. van der Leeden & F. Busing, unpublished study; also see Kreft & de Leeuw 1998). For example, in Bassiri's study, the L2/L1 sample size combination of 150/5 ($N = 750$) reached similar power to 60/25 ($N = 1{,}500$). In this study, the 30/30 rule of thumb was proposed as the minimum to reach enough power for cross-level interactions. Later, this rule of thumb was considered applicable to cross-level direct effects (Hox 2002, 2010). However, in an analytical study based on Snijders & Bosker's (1993) formulae, Scherbaum & Ferreter (2009) illustrated how the 30/30 rule was excessively demanding when the magnitude of the cross-level direct effect was medium or large, and they showed that combinations of 30/15 and 15/7, respectively, reached acceptable power. In addition, the rule was too lenient when the effect size was low. Power did not exceed 0.30 in the largest combination (40/30). In a recent simulation study, Bell et al. (2014) concluded that when effect sizes are in the small-medium range, 30 groups with 20–40 observations per group were enough to reach the conventional 0.80 statistical power for cross-level direct effects. However, for cross-level interactions, the power never exceeded 0.50. Similarly, Mathieu et al. (2012), who used the UMM when testing interactions, concluded that the power was substantially below 0.80 in most cases. In this latter study, contrary to what traditional simulation studies suggest, L1 sample size was more important than L2 sample size. On average (across slope variances, effect sizes, etc.), power was larger than 0.80 with 40 groups of 18. With smaller groups (3–7 individuals), 115 groups yielded low power (<0.40). Mathieu et al. (2012) also concluded that the power largely depended on the magnitude of the moderating effect, as well as the slope variability, and that these factors interacted with the L1 and L2 sample sizes.

Finally, for multilevel mediation, Zhang et al. (2009) showed that, when paying attention to the indirect between effect of 2–1–1 mediation by fitting the UMM, more groups with fewer observations per group led to more power than the other way around. Using the Sobel (1982) test, they concluded that a minimum of 75/8 was necessary to reach acceptable power. For those who argue that the 2–1–1 mediation should not be restricted to between effects, the conclusion was the opposite. Pituch & Stapleton (2012) showed that the power to detect the total indirect effect when using the UMM was four times lower than the power reached by using GMC(*M*). However, they used Sobel's test, which assumes that *ab* follows a normal distribution. Comparing several methods that do not assume normality, Pituch & Stapleton (2008) concluded that the bias-corrected parametric bootstrap and the empirical-*M* test showed the best performance. However, the power only reached an average of 0.80 when the size of the indirect effect was large.

Although most simulation studies have considered only two levels, some research has examined the influence of sample size in three-level designs in cluster-randomized trials, where the interest is in differences in means. Konstantopoulos (2008, 2009) and Teerenstra et al. (2008) showed that the power is higher when more L2 units per L3 unit are sampled, compared to more L1 units per L2 unit. In addition, maximizing the number of L3 units has the greatest impact on power. Moreover, the larger the clustering effect, the more important it is to increase the number of higher-level units (Konstantopoulos 2009). These recommendations are congruent with those derived from two-level designs, and they seem reasonable when the interest is in detecting cross-level direct effects. No research has focused on recommended sample size combinations to increase power for interactions in three-level models.

We agree with Tonidandel et al. (2015) that it is dangerous to rely on rigid rules of thumb about sample sizes because they just generalize to the specific simulated conditions. However, the studies reviewed do suggest that it is better to have more groups with fewer observations per group than the other way around, particularly for cross-level moderations and indirect effects (especially between indirect effects). Moreover, for cross-level interactions, additional attention must be paid to having enough individuals per group (more than 7), unless more than 100 L2 units can be collected (Mathieu et al. 2012).

One of the problems researchers encounter when making decisions about the required L1 and L2 sample sizes is that many factors have to be taken into account, especially in cross-level interactions. In fact, no simple formula exists to estimate power in this case (Scherbaum & Ferreter 2009). An interesting possibility is to estimate power based on Monte Carlo simulations. Mathieu et al. (2012) developed such a program, which can be very useful to help researchers understand how different factors (ICCs, reliability, etc.) influence power, and then decide what sample size combinations would be best considering the available information. For cross-level direct effects, the free software PINT (Power IN Two-level designs; Bosker et al. 2003) and MLPowSim (Browne et al. 2009) can also be used.

Because the lack of power is one of the major caveats in CMLM and its modified versions, including relevant covariates (at either level of analysis) may be a good option to increase power (Pituch & Stapleton 2012, Scherbaum & Ferreter 2009). In addition, Bayesian estimation methods that incorporate prior information are promising with small samples (e.g., Yuan & MacKinnon 2009, 2014). Finally, there is a research trend that defends a shift from null hypothesis testing to precise parameter estimation (see Tonidandel et al. 2015). In this case, power analysis is not as relevant because what matters is the magnitude of the population effect, which can be inferred from the confidence interval (provided that the estimate is unbiased and precise). We think these two approaches are complementary. In addition to reporting confidence intervals, it would also be a good practice to report indicators of effect size that are independent of any specific metric.

Estimating and Reporting Effect Sizes

The presence of multiple variance components in MLM complicates the estimation of effect sizes. Although global effect size indicators such as *R*-square can be estimated, local indicators, typically based on the reduction of error variances observed when comparing models with and without the predictor of interest, are more useful for understanding the importance of a particular predictor.

Among the local measures, the proportional reduction in variance components at a specific level can be used (Bryk & Raudenbush 1992). Other indicators, such as those based on OLS regression and the measures of Snijders & Boskers (1994) and Nakagawa & Schielzeth (2013), focus on total variance explained by L1 and L2 predictors (see LaHuis et al. 2014). In a simulation study, LaHuis et al. (2014) showed that these measures showed acceptable performance, except for Bryk & Raudnebush's (1992) L2 statistic. In addition, when random slopes were modeled but the cross-level interaction effects were not included in the effect size formulas, Bryk & Raudenbush's (1992) L1- and L2-specific indices performed worse than the other indices. Another possible effect size indicator is Cohen's f^2 (Cohen 1988). Selya et al. (2012) showed how to calculate f^2 when using PROC MIXED in SAS® software. This index has the advantage that it can be interpreted according to Cohen's (1988) guidelines for small, medium, and large effects, although its comparative performance with other effect-size indicators still needs to be assessed. Finally, for interventions or quasi-experiments, Hedges (2007) proposed effect size indices based on standardized mean differences, generalizing them to three-level models (Hedges 2011).

By reporting effect sizes, OPOB researchers will contribute to presenting multilevel results with rigor, relevance, and practical impact in mind (Aguinis et al. 2010). Specifically, they will increase our understanding of the importance of multilevel predictors in terms of their significance in theory and practice. For example, researchers will make more informed decisions about whether the effect of interest is important but there is not enough power to detect it. They will also enhance the comparability of results for meta-analytical studies, and they will shed light on whether their theoretical models are underspecified and more and better predictors should be proposed.

MULTILEVEL STRUCTURAL EQUATION MODELING

The use of CMLM techniques in OPOB has notably contributed to advancing knowledge in the field. However, this approach also has several limitations that applied researchers should consider. First, the CMLM approach does not generally incorporate measurement error, which results in biased parameter estimates (Bauer 2003, Li & Beretvas 2013, Lüdtke et al. 2011). Second, originally it does not distinguish and separate the between-unit and within-unit effects of L1 variables (Preacher et al. 2010). As mentioned above, the consequence of this practice is that the estimation of multilevel indirect and moderated relationships conflates effects operating at different levels, which results in biased estimations (Preacher et al. 2011, 2016). Third, it cannot model effects of L1 variables on L2 variables [bottom-up effects (Preacher et al. 2010)]. Fourth, it cannot simultaneously model all the relationships included in mediational multilevel models involving relationships between unit-level variables and cross-level relationships (Bauer 2003, Preacher et al. 2010). And fifth, it does not provide enough information to assess model fit (Bauer 2003, Preacher et al. 2010).

MSEM overcomes these limitations. In this section, we provide a brief introduction to MSEM, show how it addresses the aforementioned limitations, review the research on its performance, and discuss its advantages and limitations.

Brief Introduction to Multilevel Structural Equation Modeling

MSEM can be viewed as the integration of MLM techniques and SEM (Mehta & Neale 2005). Statistical developments of MSEM started in the late 1980s (e.g., Goldstein & McDonald 1988, McDonald & Goldstein 1989, Muthén 1989, Muthén & Satorra 1995). Since then, continuous contributions have yielded significant improvements (e.g., Ansari et al. 2002, Muthén & Asparouhov 2008, Rabe-Hesketh et al. 2004), allowing researchers to handle unbalanced group sizes, missing data, and different types of variables, as well as model random slopes (Preacher et al. 2010).

Some key features of MSEM are the following: (*a*) In two-level designs, the variance of an L1 variable (i.e., the job satisfaction reported by employees who are members of different organizations) is divided into two orthogonal latent components: the between and the within components; (*b*) constructs of interest can be modeled as latent variables with multiple indicators, which allows researchers to take measurement error into account; and (*c*) random intercepts and slopes in L1 models are considered continuous latent variables that vary across groups (Muthén & Asparouhov 2008, Lüdtke et al. 2011, Preacher et al. 2010). Some of these characteristics can be seen in the model shown in **Figure 2**. Suppose that a researcher is interested in investigating the influence of the clan organizational culture type (Quinn & Spreitzer 1991) on employees' satisfaction via employees' perceptions of participation in decision making. Imagine that the organizational culture data were provided by the CEOs of the *j*-sampled organizations, and the data on participation and job satisfaction were provided by the *i* employees sampled from each involved organization. Imagine also that the three variables were measured by means of three-item scales. In this design, organizational culture is an organization-level variable, and participation and job satisfaction are individual-level variables. **Figure 2** shows that the variance of the participation and satisfaction items ($i4_{ij}$–$i9_{ij}$ observed variables) is modeled at the within and between levels. In the within model, the individual-level factors of participation and job satisfaction are defined by multiple indicators (items). The solid circles at the end of the arrows in the within model represent random intercepts for the observed items that can vary across organizations. In the between model, these random intercepts are continuous latent variables (e.g., $i4_j$) that work as the indicators of the organizational level factors of participation and job satisfaction. The culture items are organization-level indicators with variances that can be modeled only in the between model and that define the organizational culture factor. Finally, the between model posits that the between component of participation mediates the relationships between organizational culture and the between component of satisfaction, and the within model posits that at the individual level, participation has an impact on satisfaction.

As mentioned above, one of the advantages of MSEM is that it can model measurement error by using multiple indicators for measuring latent variables. Moreover, MSEM also incorporates sampling error.

Considering Measurement and Sampling Error

Two main sources of error can be differentiated in multilevel designs (Lüdtke et al. 2011, Marsh et al. 2009). The first is measurement error involved in measuring individual and unit-level constructs. This error can be controlled for by using multiple indicators for each construct. The second is sampling error, due to the sampling of a limited number of work-unit members and the aggregation of their scores to operationalize a unit-level construct. A common practice in multilevel studies where a researcher wants to estimate the influence of a unit-level construct (e.g., work-unit climate) on an individual-level construct (e.g., job satisfaction) is to operationalize the former by aggregating the scores of the subjects who belong to the same work unit on the unit-level variable. If only a small number of subjects are sampled from each work unit, the work-unit average obtained may be an unreliable estimate of the true work-unit mean (Lüdtke et al. 2011, Marsh et al. 2009). This is shown by the ICC(2), which is commonly used to estimate the reliability of work-unit means, and whose formula shows that its value depends on the number

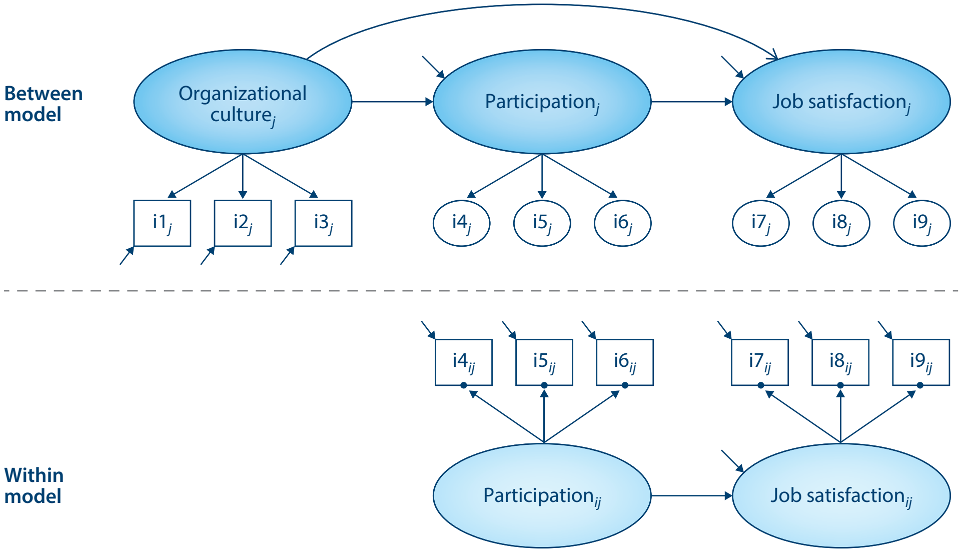

Figure 2

Representation of a two-level structural equation model based on Muthén & Muthén's (2015) rules. Observed variables are shown within squares and latent variables within ovals. Short arrows with undefined origins represent residual terms. Solid circles in the within model represent random intercepts that can vary across organizations.

of sampled subjects (Bliese 2000). Finally, an unreliable estimate of the true work-unit mean can lead to biased estimates of contextual effects.

Marsh et al. (2009) and Lüdtke et al. (2011) distinguished among different approaches to handling error in multilevel data. The uncorrected approach, which is usually implemented in CMLM studies, does not correct for either of the two sources of error mentioned above. Partial correction approaches correct for only one of the two sources of error, whereas the full correction approach corrects for both of them. Lüdtke et al. (2008, 2011) showed mathematically that the uncorrected and partial corrected approaches can yield biased estimates of contextual effects. They also conducted a series of simulation studies to determine the consequences of implementing the different approaches to estimate contextual effects. Their results showed that the full correction approach yields unbiased estimates of contextual effects "under appropriate conditions" [e.g., large number of groups (≥100) and number of subjects within groups (≥15), high ICC(1) values of the predictor variable (≥0.10), a large number of indicators (7 versus 3), and high standardized factor loadings (0.8 versus 0.6)], whereas the other approaches tended to yield biased estimates. However, they also found that the full correction approach introduces variability in the parameter estimate of contextual effects, and under certain conditions [e.g., low ICC(1), small number of groups and subjects within groups] partial correction approaches can outperform full correction in this regard.

Li & Beretvas (2013) specifically addressed the consequences of considering measurement error in MSEM in more complex models. In a simulation study, they investigated the performance of MSEM when the model involved was an upper-level multilevel mediation model (i.e., a 2–2–1 model) and the mediator and the outcome indicators showed some degree of measurement error. They compared MSEM (where the mediator and the outcome were modeled as latent variables with multiple indicators) to the CMLM approach (where the mediator and the outcome variables were operationalized as composite scores computed as the sum or mean across the indicator scores).

Their results showed that MSEM provided a higher rate of inadmissible solutions (especially when the number of groups was less than 80), but when the model converged, MSEM recovered the true indirect effect better and provided more accurate estimates. The CMLM estimates showed negative bias. Finally, the power was far below 0.80 across conditions and approaches, but it was generally slightly higher with the CMLM approach.

In summary, although MSEM generally performs better than the CMLM approach for some criteria such as parameter estimate bias, the problems reported in the reviewed studies have led some researchers (e.g., Li & Beretvas 2013) to be cautious when recommending MSEM over CMLM. Modeling measurement error in MSEM increases model complexity and the number of parameters to be estimated. To prevent convergence problems, large samples, in terms of the number of groups (>80; Li & Beretvas 2013) and the number of subjects within groups (≥15; Lüdtke et al. 2011), are needed. Additionally, these large samples will contribute to improving the power of MSEM models.

Disentangling Between-Unit and Within-Unit Effects

As mentioned above, the CMLM approach does not originally separate the between-unit and within-unit effects of L1 variables. Preacher et al. (2010, 2011, 2016) have shown that this practice leads to the estimation of multilevel indirect and moderated relationships that conflate effects operating at different levels, resulting in biased estimations.

Multilevel mediation. Within the MSEM logic, the variance of L1 variables is partitioned into between and within components. Moreover, because an L2 variable is a constant within a specific work unit, it can only explain differences in other variables (mediators or outcomes) at the work-unit level. Therefore, "any mediation effect in a model in which at least one of *X*, *M*, or *Y* is assessed at Level 2 must occur strictly at the between-group level" (Preacher et al. 2010, p. 210). Thus, when testing 2–1–1 models, unless the original multilevel model under CMLM is transformed into UMM using CWC(*M*), the indirect effect obtained by means of CMLM will conflate the between and within effects of the mediator on the outcome, and, therefore, the indirect effect estimates will be biased (Preacher et al. 2010).

However, although the UMM approach separates the two effects, it uses the group mean of the predictor and/or the mediator instead of the group's corresponding latent (true) score (Preacher et al. 2010). As mentioned above, this practice produces biased estimates of the corresponding between effect (Lüdtke et al. 2008, 2011) and, as a consequence, of the indirect effect involved (Preacher et al. 2011).

To address all these problems, Preacher et al. (2010), using Muthén & Asparouhov's (2008) approach, proposed an MSEM framework for testing multilevel mediation that integrates different mediation models, and they provided some illustrative examples. Then, in a simulation study, Preacher et al. (2011) compared the performance of MSEM to the performance of the CMLM and UMM approaches to estimate an indirect effect in a 2–1–1 model. Their results showed that MSEM outperformed the other two approaches in terms of parameter estimate bias and accuracy. Moreover, MSEM did not show convergence problems (even in the most adverse conditions), and it yielded adequate power values (>0.80) under the joint concurrence of the following conditions: number of groups ≥100, size of groups ≥20, and ICC(1)s for the mediator and the outcome variables ≥0.10. These results contrast with the problems associated with MSEM observed by Li & Beretvas (2013). The results differ likely because, whereas Li & Beretvas (2013) used multiple indicators for the mediator and the outcome, Preacher et al. (2011) used composite scores.

Multilevel moderation. In a recent article, Preacher et al. (2016) showed that because the original CMLM approach does not separate the within and between effects of L1 variables, the coefficients generally used to test for moderation in multilevel designs can conflate two or more interactions. As explained earlier, the cross-level interaction of interest ($X_{ij}W_j$) and the between-level interaction (X_jW_j) are conflated, unless the UMM is used to disentangle these two sources of variability. To illustrate this problem, imagine that a researcher wants to ascertain whether organizational culture (as measured in the example mentioned above) moderates the effect of employees' job stress on employees' tension. Because the moderator is an L2 variable and the predictor and outcome are L1 variables, this design can be represented as a 2 × (1 → 1) model (Preacher et al. 2016). If we decompose job stress and tension into their within and between components, and we realize that job stress can have a within and a between effect on tension (as shown in **Figure 2** for the participation-satisfaction relationship), it is easy to see that organizational culture can moderate the within effect of job stress on tension (a cross-level interaction, $X_{ij}W_j$) and also the "Stress → Tension" between effect operating at the organizational level (an L2 interaction, X_jW_j). The CMLM approach yields a single coefficient to estimate the expected cross-level interaction, but it conflates the two aforementioned interactions (Preacher et al. 2016). These researchers offer an MSEM framework that solves this problem in different multilevel designs, and they show how to implement it. Moreover, they conducted a simulation study whose results indicated that the MSEM method investigated to test for multilevel moderation (latent moderation structural equations, LMS) did not show convergence problems and generally yielded unbiased interaction estimates. However, its power was low (<0.80).

Modeling Bottom-Up Effects

The CMLM approach assumes that outcome variables reside at lower levels of analysis and cannot affect higher-level variables. Thus, relationships in a multilevel system involve variables influencing other variables at the same or lower levels (Krull & MacKinnon 2001, Mathieu & Taylor 2007). This precludes the analysis of bottom-up effects where an L1 variable affects an L2 variable (also known as upward influence and micro-macro effects). However, relationships of this type have been suggested in the past in the field (Griffin 1997, Schneider 1987). For instance, individual helping behavior can contribute to building a work team climate of support. Some methods have been proposed to estimate bottom-up effects (Griffin 1997, Croon & van Veldhoven 2007), but they involve different steps that make the procedures fairly unfriendly.

MSEM can be used to test models in which outcome and/or mediator variables are measured at higher levels and predictors at lower ones (e.g., 1–1–2, 1–2–1; Preacher et al. 2010). The key point is to realize that, as mentioned above, if a relationship involves an L2 variable, then it can only exist at the between level. Imagine that a researcher wants to investigate the influence of employees' perceived participation on organizational performance via employee satisfaction. Suppose that participation and satisfaction are measured as in the example represented in **Figure 2**, and organizational performance is operationalized as sales growth (i.e., a 1–1–2 model). It is the between component of participation that impacts the between component of satisfaction, which in turn impacts organizational performance. Thus, within MSEM, upward influences are modeled as between effects involving the between components of the corresponding L1 variables.

Simultaneous Testing

In many instances, the CMLM approach cannot directly and simultaneously test all the relationships involved in a multilevel model (e.g., a multilevel mediation model). In these cases, the traditional approach is to combine information from several methods. For example, to test the indirect effect (*ab*) involved in a 2–2–1 model, the *a* coefficient estimating the 2–2 relationship is obtained by using OLS regression at L2, whereas the *b* coefficient estimating the 2–1 relationship is obtained by implementing CMLM techniques (e.g., Chen et al. 2007, Mathieu & Taylor 2007). MSEM allows researchers to estimate all the involved relationships directly and simultaneously, a feature that becomes more important as model complexity increases (Preacher et al. 2010).

Assessing Fit

In the CMLM approach, the goodness-of-fit assessment of a hypothesized model is problematic because "there is no logical saturated model with which to compare a particular fitted model," and, consequently, there is no single inferential test to assess model fit (Curran 2003, p. 564). By contrast, there are several fit indices in the SEM literature that can be used to assess MSEM models. However, there are some problems associated with applying the SEM approach used in single-level models to assess the fit of the entire MSEM model (i.e., the standard approach). The test of exact fit in SEM tests the hypothesis that the population covariance matrix (Σ) equals the covariance matrix reproduced by the hypothesized model and its parameters [($\Sigma\theta$); that is, H_0: $\Sigma = \Sigma\theta$]. The standard approach in MSEM tests the joint hypothesis that the L1 (within) and L2 (between) population covariance matrices (Σ_W and Σ_B, respectively) equal the corresponding L1 and L2 covariance matrices reproduced by the hypothesized model [$\Sigma_w(\theta)$ and $\Sigma_B(\theta)$, respectively]; that is, H_0: $\Sigma_W = \Sigma_W$ (θ) and $\Sigma_B = \Sigma_B$ (θ) (Ryu 2014). This approach has the following problems (Ryu 2014, Ryu & West 2009, Yuan & Bentler 2007). First, because the MSEM model is evaluated simultaneously at both levels and the sample size at the lower level (e.g., individual) is generally much larger than the sample size at the higher level (e.g., group), model fit assessment is likely to be dominated by model fit at the lower level. Second, when poor model fit is observed, it is not clear where its cause(s) resides: at L1, at L2, or at both levels. These problems have led to some alternatives aimed at developing level-specific methods to assess model fit in MSEM models.

Yuan & Bentler's (2007) segregating procedure is based on the idea of separating a two-level model into two single-level models whose fit is assessed independently at the corresponding level. Ryu & West (2009) proposed using partially saturated models to obtain level-specific fit indices. In both procedures, once the test of exact fit for each level-specific model is obtained, other fit indices derived from it (e.g., CFI, RMSEA) can be computed. The simulation studies conducted to evaluate the performance of these procedures showed that (*a*) the level-specific methods successfully detected a poor-fitting model at L2, whereas the standard approach did not; (*b*) both approaches detected a poor-fitting model at L1; (*c*) the performance of level-specific maximum likelihood (ML) test statistics was affected by skewness and kurtosis; and (*d*) the partially saturated models were able to identify the specific level at which poor fit occurred in models with latent interactions at both levels (Ryu 2011, Ryu & West 2009, Schermelleh-Engel et al. 2014, Yuan & Bentler 2007). Overall, these results suggest that researchers should use level-specific methods and indices to assess the fit of MSEM models.

Multilevel Structural Equation Modeling: Advantages and Limitations

The previous pages show that MSEM has some strong advantages because it solves important issues that CMLM techniques cannot address or cannot handle in a proper manner. Moreover, MSEM models can include multiple mediators and moderated mediation (Preacher et al. 2010, 2016). However, MSEM also presents problems and limitations. MSEM models tend to be complex, especially when they include multiple indicators per focal variable. Complex models may not perform well with modest sample sizes, showing problems of nonconvergence and variability in parameter estimates (Li & Beretvas 2013, Lüdtke et al. 2011). These problems have motivated research specifically focused on sample size requirements and estimation methods in an effort to determine the conditions under which MSEM models work best.

Since Hox & Maas' (2001) investigation on the performance of a multilevel confirmatory factor analysis model, several studies have examined the effect of sample size on the functioning of MSEM models with structural relations among latent variables. This issue is especially important in cross-cultural research [e.g., the GLOBE study (House et al. 2004)], where the number of higher-level units (countries) is generally low. In this context, Cheung & Au (2005), keeping the number of countries constant ($N = 27$), observed that the individual-level results were quite stable, even with small (50) within-unit samples. However, increasing the latter did not necessarily improve parameter estimation at the higher level. Meuleman & Billiet (2009), in a series of simulation studies, assessed the performance of MSEM with ML robust estimation methods with varying sizes of the higher-level sample (20 to 100) when the size of the within-unit samples was large. They found that (*a*) the estimation accuracy of the between-model parameters increased with the number of higher-level units; (*b*) model complexity had an important influence on estimation accuracy; and (*c*) for simple between-level models, detecting a large between effect required 60 units at least; however, detecting smaller effects required more than 100 units. Continuing this line of research, Hox et al. (2012) reanalyzed Meuleman & Billiet's (2009) data using Bayesian estimation. They observed that a sample of 20 higher-level units was sufficient to obtain accurate estimates of the factor loadings and the structural parameters in the between model. However, the power was generally low (except when effect size was very large). Recently, Hox et al. (2014) conducted a simulation study to ascertain the lowest number of higher-level units needed when MSEM is used to estimate an indirect between effect in a three-path mediational model (intervention-attitude-intention-behavior). They compared Bayesian and ML estimation and concluded that the former worked better than the latter when the number of higher-level units was small (25). These results suggest that Bayesian estimation can be a reasonable option when the number of higher-level units is small (Lüdtke et al. 2011, Marsh et al. 2009).

Finally, although MSEM models can be extended to three-level data (see Preacher 2011), we did not find any simulation study that investigated the performance of three-level MSEM under varied conditions. Considering that three-level data may be more available in the near future thanks to the big data movement, studies filling this gap in the literature are certainly needed.

PRACTICAL GUIDELINES FOR ORGANIZATIONAL PSYCHOLOGY AND ORGANIZATIONAL BEHAVIOR RESEARCHERS

On the basis of the literature reviewed above, we next offer several practical guidelines for OPOB researchers needing multilevel techniques:

1. After considering the advantages and limitations of both approaches, we recommend MSEM over CMLM and its modifications (e.g., UMM), especially for testing models that include indirect effects and moderated relationships.
2. To test MSEM models that include multiple indicators with ML estimation, a sample size of at least 100 work units (Li & Beretvas 2013, Lüdtke et al. 2011, Meuleman & Billiet 2009) and 15 subjects per unit (Lüdtke et al. 2011) is advisable. If actual sample sizes fall well below these figures, researchers should consider using Bayesian estimation methods (Hox et al. 2012, 2014).
3. Model complexity affects the performance of MSEM models (Meuleman & Billiet 2009). If nonconvergence appears, researchers should consider simplifying their models. This can

be done by imposing invariance constraints across levels on some parameters [e.g., loadings (Lüdtke et al. 2011)], or by fitting partial correction models with composite scores (Marsh et al. 2009).

4. To assess model fit, level-specific tests and indices are recommended (Ryu & West 2009, Yuan & Bentler 2007).

If researchers cannot test their multilevel hypotheses with MSEM and have to use CMLM, we propose the following guidelines:

5. Predictors should be rescaled to have a meaningful interpretation of regression coefficients. For the L1 predictors, rescaling should consider the theoretical process that underlies the focal phenomenon (e.g., group comparison versus absolute standing on a construct) before deciding which centering option (CWC or GMC) is best (Aguinis et al. 2013, Enders 2013). Regardless of the choice, adding the aggregate L1 predictor at L2 can solve the problem of conflated variance and provide information about whether a predictor is meaningful at both levels of analysis to fully describe the relationship of interest (Enders 2013).
6. If researchers do not want to test or control for contextual effects, then CWC should be the choice for testing cross-level interactions (Aguinis et al. 2013). In this way, the problem of conflated variance is still avoided.
7. If both the between and within effects are relevant, introducing the means of the L1 predictors as an L2 covariate can increase power and improve model specification (Pituch & Stapleton 2012, Raudenbush 1997, Scherbaum & Ferreter 2009).
8. Assuming a medium effect size, when testing cross-level direct effects, researchers should aim for a minimum of 30 groups of between 15–20 individuals (e.g., Bell et al. 2014, Scherbaum & Ferreter 2009). For cross-level interactions, the 40/18 combination is suggested to reach acceptable power (Mathieu et al. 2012). Finally, for multilevel mediation, combinations may depend on the type of mediation. For 2–1–1 models, 75/8 is recommended to test between-group indirect effects (Zhang et al. 2009). When it is reasonable to estimate the indirect effect as the sum of the between- and the cross-level indirect effects, 40/20 would be enough for large effects. For medium effects, more or larger groups should be sampled (Pituch & Stapleton 2012). Although helpful, these recommendations should be interpreted cautiously; they stem from simulation conditions that may not be generalizable to a specific study. Power analysis using available programs (Bosker et al. 2003, Mathieu et al. 2012) is strongly advised.
9. Regardless of the approach implemented (CMLM or MSEM), OPOB researchers should systematically report confidence intervals, effect sizes, and power when testing multilevel hypotheses. Thus, they will provide richer information about the relationships observed.

As a complement to these guidelines, **Table 1** provides a set of questions and responses with the aim of helping OPOB researchers make informed decisions when designing their multilevel studies. As we mentioned above, it is dangerous to rely on rigid rules of thumb about certain design features (e.g., sample size) because the results of the reviewed simulation studies can only be generalized to the specific simulated conditions (Tonidandel et al. 2015).

OPPORTUNITIES FOR FUTURE RESEARCH

Multilevel modeling can contribute to theoretical advancement and research development in OPOB. By offering ways to investigate relationships between constructs that reside at different levels, this method promotes a more comprehensive understanding of organizational phenomena, and it can help to discover less obvious predictors, interactions, and outcomes (Hackman 2003).

Table 1 Questions and responses to help organizational psychology and organizational behavior (OPOB) researchers to make informed decisions when designing multilevel studies

Questions	Responses	Relevant references	Related tools
General			
When does the practice of not using multilevel techniques to analyze nested data become problematic?	This is problematic when the ICC(1) values of the involved variables are greater than 0.05. However, when ICC(1) ≤ 0.05, the consequences of ignoring the nested structure of the data are negligible.	Finch & French 2011, Julian 2001	**https://cran.r-project.org/web/packages/ICC/index.html**
CMLM			
What are the recommended methods to ascertain whether there is enough variability across intercepts to test for cross-level direct effects?	Because the SEs of the intercept residual variances are inaccurate, and null-hypothesis tests for these residual variances have low power, the recommendation is to quantify the proportion of criterion variance attributed to intercept differences across groups by means of ICC(1). Values of ∼0.05 in ICC(1) may be considered large enough.	Heck et al. 2013, LeBreton & Senter 2008	**https://cran.r-project.org/web/packages/ICC/index.html**
What are the recommended methods to ascertain whether there is enough variability across slopes to test for cross-level interactions?	Because the SEs of the slope residual variances are inaccurate, and null-hypothesis tests these residual variances have low power, the recommendation is to quantify the proportion of criterion variance attributed to slope differences across groups by means of ICC(β). Values of ∼0.05 may be considered large enough.	Aguinis & Culpepper 2015	**https://cran.r-project.org/web/packages/iccbeta/index.html**
What are the recommended options for centering predictors and moderators in MLM?	For L2 predictors and moderators GMC is the only option. For L1 predictors and moderators, CWC(M) and GMC(M) are valid options to adequately partition the between-group and the within-group variance. The choice between CWC(M) or GMC(M) should be based on whether individuals' relative position within a group (i.e., the frog-pond effect) or their absolute position, respectively, is considered the relevant predictor. If researchers have reasons to exclude contextual effects, CWC should be the choice to obtain an accurate estimate of within-group slopes. An exception to this recommendation is the case of cluster-randomized designs: When L1 variables are not of interest and researchers just want to control for them as covariates, GMC should be the choice.	Bryk & Raudenbush 1992, Enders 2013, Enders & Tofighi 2007, Hofmann & Gavin 1998, Kreft et al. 1995, Dalal & Zickar 2012, Paccagnella 2006, Zhang et al. 2009	None

(*Continued*)

Table 1 ***(Continued)***

Questions	Responses	Relevant references	Related tools
What are the necessary sample sizes at different levels of analysis to obtain unbiased estimates of cross-level regression coefficients, and accurate SEs?	Small sample sizes such as 20 groups of 10 individuals, and even 10 groups of 5, have been reported to be large enough to obtain unbiased estimates. However, to get accurate SEs, larger numbers of L2 units (a minimum of 30) have been found to be necessary. For smaller numbers of L2 units, Bayesian estimation methods are promising options.	Bell et al. (2014), Maas & Hox (2000, 2004b, 2005), Mok (1995), Stegmueller (2013), Van der Leeden et al. (1997)	None
What are the necessary sample sizes at different levels of analysis to achieve enough power when testing for cross-level direct effects and interactions?	It is better to have more groups with fewer individuals per group than the other way around, particularly for cross-level moderations. When effect sizes are in the small–medium range, 30 groups with 20–40 individuals per group have been reported to be enough to reach a statistical power of 0.80 for cross-level direct effects. For medium and large effects, 30 groups of 15 and 15 groups of 7, respectively, have been reported to be enough. For cross-level interactions, more L2 units are necessary (on average across conditions: sample size combinations of 40/18). Special attention must be paid to having enough individuals per group (8 or more), unless more than 100 L2 units can be collected. Power analysis should be systematically carried out when designing the studies to make an informed decision about the necessary sample sizes at different levels of analysis.	Bassiri (1988), Bell et al. (2014), Kim (1990), Mathieu et al. (2012), Scherbaum & Ferreter (2009)	**https://www.stats.ox.ac.uk/~snijders/multilevel.htm/**, **http://www.bristol.ac.uk/cmm/software/mlpowsim/**, **http://www.hermanaguinis.com/crosslevel.html**
What are the best conditions (sample sizes and methods) to test for multilevel mediation?	Methods that take into account that the indirect effect, *ab*, is not normally distributed (such as bootstrap or the empirical-M test) should be used. Indirect effects and SEs are typically accurate, even for small sample size combinations. Regarding power, for indirect between effects, it is better to have more groups with fewer individuals per group than the other way around.	Bauer et al. (2006), Krull & MacKinnon (1999, 2001), Zhang et al. (2009)	**http://quantpsy.org/medmc/medmc111.htm**, **https://cran.r-project.org/web/packages/mediation/index.html**

(*Continued*)

Table 1 *(Continued)*

Questions	Responses	Relevant references	Related tools
What are the recommended measures of effect size?	Measures that focus on total variance explained by L1 and L2 predictors using a hierarchical approach: Snijders & Bosker's (1994) measure, the OLS-based approach, and Nakagawa & Schielzeth's (2013) measure, show acceptable performance in terms of bias, consistency, and efficiency. Among them, Snijders & Bosker's is one of the least biased for random intercept and random slope models; it is easy to calculate and has an intuitive interpretation.	LaHuis et al. 2014	**http://psych-scholar.wright.edu/lahuis/software/multileveleffect-size-function**
MSEM			
Under which conditions does the full correction approach in MSEM with ML estimation work best?	The best conditions are a sample size of 100 work units and 15 subjects per unit, ICC(1) values greater than 0.10, a large number of measurement indicators (7), and high standardized factor loadings (0.80). However, power is generally below 0.80.	Li & Beretvas 2013, Lüdtke et al. 2011, Meuleman & Billiet 2009	None
Considering these demanding conditions and the associated problems (e.g., low power), which alternatives seem reasonable?	Within ML estimation, using composite scores as single indicators of constructs instead of multiple indicators (e.g., items) (that is, fitting partial correction MSEM models) offers a reasonable alternative. With conditions such as number of groups ≥100, size of groups ≥20, and ICC(1) values ≥0.10, this alternative yields adequate estimates for testing indirect effects in "2–1–1" models, and it provides adequate power values (>0.80). Using Bayesian estimation methods is another alternative. These methods work better than ML when the number of sampled higher-level units is as small as 25.	Marsh et al. 2009; Preacher et al. 2011; Hox et al. 2012, 2014	None
Regarding the assessment of model fit, what are the recommended methods?	Level-specific methods and indices are recommended. They indicate the investigated model's fit at the within and between levels, making it easier to identify the causes of potential misfit.	Ryu 2011, Ryu & West 2009, Schermelleh-Engel et al. 2014, Yuan & Bentler 2007	**http://www3.nd.edu/~kyuan/multilevel/Multi-Single.sas**

Abbreviations: CMLM, conventional multilevel modeling; CWC, centering within clusters; GMC, grand mean centering; CWC(M), centering within clusters reintroducing the means of the involved predictors at L2; GMC(M), grand mean centering reintroducing the means of the involved predictors at L2; ICC(1), intraclass correlation coefficient 1; ML, maximum likelihood; MSEM, multilevel structural equation modeling; OLS, ordinary least squares; SE, standard errors; SAS, Statistical Analysis System.

This is especially important in topics traditionally dominated by a single-level perspective [e.g., job stress (Peiró 2008)]. Current multilevel methods provide OPOB researchers with great opportunities for theoretical advancement and research development in the following areas: modeling bottom-up effects, investigating level-specific interactions, and testing homologies.

The widely extended idea that multilevel modeling requires the outcomes to reside at the lower level has precluded research about bottom-up effects. However, different authors have suggested and investigated these relationships (Chen 2005, Griffin 1997, Neal & Griffin 2006). The key issue to remember is that the between component of the lower-level predictor is the one involved in the relationship. Therefore, when theorizing about bottom-up effects, researchers must first clarify the meaning of the between component, considering its similarities and differences with its within counterpart. Then, models that include bottom-up effects or a combination of bottom-up and top-down effects can be tested by using MSEM (see Preacher et al. 2010).

As shown above, the decomposition of lower-level variables into their within and between components allows researchers to see that cross-level interactions estimated by using CMLM techniques may combine different interaction effects (Preacher et al. 2016). Thus, when investigating interactions within a multilevel framework, the recommendation is to decompose lower-level variables, establish their within and between effects, and identify the different interaction effects that may be operating. By doing so, it may be possible to uncover interactions that otherwise would have gone unnoticed. This "discovery" may stimulate theoretical development in an effort to explain "previously unseen" interactions. Generally, investigating specific interactions that involve specific components of lower-level variables can lead researchers to formulate more fine-grained arguments and hypotheses.

One of the interests of multilevel research on OPOB is to ascertain whether relationships among constructs can be generalized across levels. This involves the testing of homologies (i.e., theories or models positing parallel relationships across levels). Empirical research on homologies has been relatively scant in OPOB. For some time, this was probably due to the limitations of the available methods (Chen et al. 2005). Current MSEM methods allow researchers to directly test homologies (see Guenole 2016, Huhtala et al. 2015). The use of these methods can help researchers to advance scientific knowledge about relationships that generalize across levels.

CONCLUSION

"Complex substantive issues require sophisticated methodologies" (Marsh et al. 2009, p. 765). The understanding of these methodologies stimulates new questions that can foster the formulation of new theories and hypotheses. Through this article, we hope to have contributed to improving OPOB researchers' understanding of current multilevel modeling methods, their advantages and limitations, the conditions for their use, and the opportunities they offer.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We want to thank Herman Aguinis, Fred Morgeson, Jeremy Dawson, M. Gloria González-Morales, and Inés Tomás for their constructive comments on an earlier version of this review.

LITERATURE CITED

Comprehensive review of factors to consider when testing and interpreting cross-level interactions.

Aguinis H, Culpepper SA. 2015. An expanded decision-making procedure for examining cross-level interaction effects with multilevel modeling. *Organ. Res. Methods* 18:155–76

Aguinis H, Gottfredson RK, Culpepper SA. 2013. Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *J. Manag*. 39:1490–528

Aguinis H, Werner S, Abbott JL, Angert C, Park JH, Kohlhausen D. 2010. Customer-centric science: reporting significant research results with rigor, relevance, and practical impact in mind. *Organ. Res. Methods* 13:515–39

Ansari A, Jedidi K, Dube L. 2002. Heterogeneous factor analysis models: a Bayesian approach. *Psychometrika* 67:49–78

Bassiri D. 1988. *Large and small sample properties of maximum likelihood estimates for hierarchical linear models*. PhD Thesis, Mich. State Univ., East Lansing

Bauer DJ. 2003. Estimating multilevel linear models as structural equation models. *J. Educ. Behav. Stat.* 28:135–67

Bauer DJ, Curran PJ. 2005. Probing interactions in fixed and multilevel regression: inferential and graphical techniques. *Multivar. Behav. Res.* 40:373–400

Bauer DJ, Preacher KJ, Gil KM. 2006. Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: new procedures and recommendations. *Psychol. Methods* 11:142–63

Bell BA, Morgan GB, Schoeneberger JA, Kromrey JD, Ferron JM. 2014. How low can you go? *Methodology* 1:86–92

Berkhof J, Snijders TA. 2001. Variance component testing in multilevel models. *J. Educ. Behav. Stat.* 26:133–52

Bliese PD. 2000. Within-group agreement, non-independence, and reliability: implications for data aggregation and analysis. In *Multilevel Theory, Research, and Methods in Organizations*, ed. KJ Klein, SWJ Kozlowski, pp. 349–81. San Francisco: Jossey-Bass

Bliese PD, Hanges PJ. 2004. Being both too liberal and too conservative: the perils of treating grouped data as though they were independent. *Organ. Res. Methods* 7:400–17

Bosker RJ, Snijders TAB, 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)*. Groningen, Neth.: Neth. Org. Sci. Res. **https://www.stats.ox.ac.uk/~snijders/Pint21_UsersManual.pdf**

Browne WJ, Lahi MG, Parker RMA. 2009. *A Guide to Sample Size Calculations for Random Effect Models via Simulation and the MLPowSim Software Package*. Bristol, UK: Univ. Bristol

Bryk AS, Raudenbush SW. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage

Burstein L, Linn RL, Capell FJ. 1978. Analyzing multilevel data in the presence of heterogeneous within-class regressions. *J. Educ. Stat.* 3:347–83

Chen G. 2005. Newcomer adaptation in teams: multilevel antecedents and outcomes. *Acad. Manag. J.* 48:101–16

Chen G, Bliese PD, Mathieu JE. 2005. Conceptual framework and statistical procedures for delineating and testing multilevel theories of homology. *Organ. Res. Methods* 8:375–409

Chen G, Kirkman BL, Kanfer R, Allen D, Rosen B. 2007. A multilevel study of leadership, empowerment, and performance in teams. *J. Appl. Psychol.* 92:331–46

Cheung MWL, Au K. 2005. Applications of multilevel structural equation modeling to cross-cultural research. *Struct. Equ. Model.* 12:598–619

Cohen JE. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Assoc., Inc.

Croon MA, van Veldhoven JPM. 2007. Predicting group-level outcome variables from variables measured at the individual level: a latent variable multilevel model. *Psychol. Methods* 12:45–57

Curran PJ. 2003. Have multilevel models been structural equation models all along? *Multivar. Behav. Res.* 38:529–69

Dalal DK, Zickar MJ. 2012. Some common myths about centering predictor variables in moderated multiple regression and polynomial regression. *Organ. Res. Methods* 15:339–62 de Leeuw J, Kreft I. 1986. Random coefficient models for multilevel analysis. *J. Educ. Stat.* 11:57–85

Enders C. 2013. Centering predictors and contextual effects. In *The SAGE Handbook of Multilevel Modeling*, ed. MA Scott, JS Simonoff, BD Marx, pp. 89–108. London: Sage

Enders CK, Tofighi D. 2007. Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychol. Methods* 12:121–38

Detailed overview and practical recommendations about centering options, with and without contextual effects.

Finch WH, French BF. 2011. Estimation of MIMIC model parameters with multilevel data. *Struct. Equ. Model.* 18:229–52

Gavin MB, Hofmann DA. 2002. Using hierarchical linear modeling to investigate the moderating influence of leadership climate. *Leadersh. Quart.* 13:15–33

Goldstein H. 1986. Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* 73:43–56

Goldstein H, McDonald RP. 1988. A general model for the analysis of multilevel data. *Psychometrika* 53:455–67

González-Romá V, Hernández A. 2014. Climate uniformity: its influence on team communication quality, task conflict, and team performance. *J. Appl. Psychol.* 99:1042–58

Griffin MA. 1997. Interaction between individuals and situations: using HLM procedures to estimate reciprocal relationships. *J. Manag.* 23:759–73

Guenole N. 2016. The importance of isomorphism for conclusions about homology: a Bayesian multilevel structural equation modeling approach with ordinal indicators. *Front. Psychol.* 7(289):1–17

Hackman JR. 2003. Learning more by crossing levels: evidence from airplanes, hospitals, and orchestras. *J. Organ. Behav.* 24:905–22

Heck RH, Thomas SL, Tabata LN. 2013. *Multilevel and Longitudinal Modeling with IBM SPSS*. New York: Routledge

Heck RH, Thomas SL. 2015. *An Introduction to Multilevel Modeling Techniques: MLM and SEM Approaches Using Mplus*. New York: Routledge

An easy-to-follow introduction to CMLM and MSEM.

Hedges LV. 2007. Effect sizes in cluster randomized designs. *J. Educ. Behav. Stat.* 32:341–70

Hedges LV. 2011. Effect sizes in three-level cluster-randomized experiments. *J. Educ. Behav. Stat.* 36:346–80

Hitt MA, Beamish PW, Jackson SE, Mathieu JE. 2007. Building theoretical and empirical bridges across levels: multilevel research in management. *Acad. Manag. Rev.* 50:1385–99

Hofmann DA. 1997. An overview of the logic and rationale of hierarchical linear models. *J. Manag.* 23:723–44

Introductory tutorial to CMLM, with an overview of typical series of models to be investigated.

Hofmann DA, Gavin MB. 1998. Centering decisions in hierarchical linear models: implications for research in organizations. *J. Manag.* 24:623–41

Hofmann DA, Morgeson FP, Gerras SJ. 2003. Climate as a moderator of the relationship between leader-member exchange and content specific citizenship: safety climate as an exemplar. *J. Appl. Psychol.* 88:170–78

House R, Rousseau DM, Thomas-Hunt M. 1995. The meso paradigm: a framework for the integration of micro and macro organizational behavior. *Res. Organ. Behav.* 17:71–114

House RJ, Hanges PJ, Javidan M, Dorfman PW, Gupta V, eds. 2004. *Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies*. Thousand Oaks, CA: Sage

Hox JJ. 2002. *Multilevel Analysis*. Mahwah, NJ: Lawrence Earlbaum

Hox JJ. 2010. *Multilevel Analysis: Techniques and Applications*. New York: Routledge. 2nd ed.

Hox JJ, Maas CJ. 2001. The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Struct. Equ. Model.* 8:157–74

Hox JJ, Moerbeek M, Kluytmans A, van de Schoot R. 2014. Analyzing indirect effects in cluster randomized trials. The effect of estimation method, number of groups and group sizes on accuracy and power. *Front. Psychol.* 5(78):1–7

Hox JJCM, van de Schoot R, Matthijsse S. 2012. How few countries will do? Comparative survey analysis from a Bayesian perspective. *Surv. Res. Methods* 6:87–93

Huhtala M, Tolvanen A, Mauno S, Feldt T. 2015. The associations between ethical organizational culture, burnout, and engagement: a multilevel study. *J. Bus. Psychol.* 30:399–414

Julian MW. 2001. The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Struct. Equ. Model.* 8:325–52

Kenny DA, Korchmaros JD, Bolger N. 2003. Lower level mediation in multilevel models. *Psychol. Methods* 8:115–28

Kim KS. 1990. *Multilevel data analysis: a comparison of analytical alternatives*. PhD Thesis, Univ. Calif., Los Angeles

Klein KJ, Dansereau F, Hall RJ. 1994. Levels issues in theory development, data-collection, and analysis. *Acad. Manag. Rev.* 19:195–229

Konstantopoulos S. 2008. The power of the test for treatment effects in three-level cluster randomized designs. *J. Res. Educ. Eff.* 1:66–88

Konstantopoulos S. 2009. Incorporating cost in power analysis for three-level cluster- randomized designs. *Eval. Rev.* 33:335–57

Kozlowski SWJ, Klein KJ. 2000. A multilevel approach to theory and research in organizations. Contextual, temporal, and emergent processes. In *Multilevel Theory, Research, and Methods in Organizations*, ed. KJ Klein, SWJ Kozlowski, pp. 3–90. San Francisco: Jossey-Bass

Kreft I, de Leeuw J. 1998. *Introducing Multilevel Modeling*. London: Sage

Kreft IGG, de Leeuw J, Aiken LS. 1995. The effect of different forms of centering in hierarchical linear models. *Multivar. Behav. Res.* 30:1–21

Krull JL, MacKinnon DP. 1999. Multilevel mediation modeling in group-based intervention studies. *Eval. Rev.* 23:418–44

Krull JL, MacKinnon DP. 2001. Multilevel modeling of individual and group level mediated effects. *Multivar. Behav. Res.* 36:249–77

Lance CE, Vandenberg RJ, ed. 2015. *More Statistical and Methodological Myths and Urban Legends*. New York: Routledge

LaHuis DM, Ferguson MW. 2009. The accuracy of significance tests for slope variance components in multilevel random coefficient models. *Organ. Res. Methods* 12:418–35

LaHuis DM, Hartman MJ, Hakoyama S, Clark PC. 2014. Explained variance measures for multilevel models. *Organ. Res. Methods* 17:433–51

LeBreton JM, Senter JL. 2008. Answers to twenty questions about interrater reliability and interrater agreement. *Organ. Res. Methods* 11:815–52

Lee S, Dalal RS. 2016. Climate as situational strength: Safety climate strength as a cross-level moderator of the relationship between conscientiousness and safety behaviour. *Eur. J. Work Organ. Psy.* 25:120–32

Li X, Beretvas SN. 2013. Sample size limits for estimating upper level mediation models using multilevel SEM. *Struct. Equ. Model.* 20:241–64

LoPilato AC, Vandenberg RJ. 2015. The not-so-direct cross-level direct effect. See Lance & Vandenberg 2015, pp. 292–310

Lüdtke O, Marsh HW, Robitzsch A, Trautwein U. 2011. A 2 × 2 taxonomy of multilevel latent contextual models: accuracy–bias trade-offs in full and partial error correction models. *Psychol. Methods* 16:444–67

Shows how measurement and sampling error affect estimates of contextual effects and provides syntax.

Lüdtke O, Marsh HW, Robitzsch A, Trautwein U, Asparouhov T, Muthén B. 2008. The multilevel latent covariate model: a new, more reliable approach to group-level effects in contextual studies. *Psychol. Methods* 13:203–29

Lüdtke O, Robitzsch A, Trautwein U, Kunter M. 2009. Assessing the impact of learning environments: how to use student ratings of classroom or school characteristics in multilevel modeling. *Contemp. Educ. Psychol.* 34:120–31

Maas CJ, Hox JJ. 2000. *Robustness of multilevel parameter estimates against small sample sizes*. Presented at Int. Conf. Log. Methodol., 5th, Cologne, Germany

Maas CJ, Hox JJ. 2004a. The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Comput. Stat. Data An.* 46:427–40

Maas CJ, Hox JJ. 2004b. Robustness issues in multilevel regression analysis. *Stat. Neerl.* 58:127–37

Maas CJ, Hox JJ. 2005. Sufficient sample sizes for multilevel modeling. *Methodology* 1:86–92

MacKinnon DP, Fairchild AJ, Fritz MS. 2007. Mediation analysis. *Annu. Rev. Psychol.* 58:593–614

Marsh HW, Lüdtke O, Robitzsch A, Trautwein U, Asparouhov T, et al. 2009. Doubly-latent models of school contextual effects: integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivar. Behav. Res.* 44:764–802

Mathieu JE, Aguinis H, Culpepper SA, Chen G. 2012. Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *J. Appl. Psychol.* 97:951–66

Mathieu JE, Chen G. 2011. The etiology of the multilevel paradigm in management research. *J. Manag.* 37:610–41

Mathieu JE, Taylor SR. 2007. A framework for testing meso-mediational relationships in Organizational Behavior. *J. Organ. Behav.* 28:141–72

McDonald RP, Goldstein H. 1989. Balanced versus unbalanced designs for linear structural relations in two-level data. *Br. J. Math. Stat. Psychol.* 42:215–32

Mehta PD, Neale MC. 2005. People are variables too: multilevel structural equations modeling. *Psychol. Methods* 10:259–84

Meuleman B, Billiet J. 2009. A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM? *Surv. Res. Methods* 3:45–58

Mok M. 1995. Sample size requirements for 2-level designs in educational research. *Multilevel Modelling Newsletter*, Vol. 7(2), June. **http://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/new7-2.pdf**

Moerbeek M. 2004. The consequence of ignoring a level of nesting in multilevel analysis. *Multivar. Behav. Res.* 39:129–49

Muthén B. 1989. Latent variable modeling in heterogeneous populations. *Psychometrika* 54:557–85

Muthén B, Asparouhov T. 2008. Growth mixture modeling: analysis with non-Gaussian random effects. In *Longitudinal Data Analysis*, ed. G Fitzmaurice, M Davidian, G Verbeke, G Molenberghs, pp. 143–65. Boca Raton, FL: Chapman & Hall/CRC

Muthén B, Satorra A. 1995. Complex sample data in structural equation modeling. *Sociol. Methodol.* 25:267–316

Muthén LK, Muthén B. 2015. *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén

Nakagawa S, Schielzeth H. 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods Ecol. Evol.* 4:133–42

Naumann SE, Bennett N. 2000. A case for procedural justice climate: Development and test of a multilevel model. *Acad. Manag. J.* 43:881-889.

Erratum

Neal A, Griffin MA. 2006. A study of the lagged relationships among safety climate, safety motivation, safety behavior, and accidents at the individual and group levels. *J. Appl. Psychol.* 1:946–53

Paccagnella O. 2006. Centering or not centering in multilevel models? The role of the group mean and the assessment of group effects. *Eval. Rev.* 30:66–85

Peiró JM. 2008. Stress and coping at work: new research trends and their implications for practice. In *The Individual in the Changing Working Life*, ed. K Näswall, J Hellgren, M Sverke. pp. 284–310. New York: Cambridge Univ. Press

Pituch KA, Stapleton LM. 2008. The performance of methods to test upper-level mediation in the presence of nonnormal data. *Multivar. Behav. Res.* 43:237–67

Pituch KA, Stapleton LM. 2012. Distinguishing between cross-and cluster-level mediation processes in the cluster randomized trial. *Sociol. Method. Res.* 41:630–70

Preacher KJ. 2011. Multilevel SEM strategies for evaluating mediation in three-level data. *Multivar. Behav. Res.* 43:691–731

Preacher KJ, Curran PJ, Bauer DJ. 2006. Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *J. Educ. Behav. Stat.* 31:437–48

Preacher KJ, Zhang Z, Zyphur MJ. 2011. Alternative methods for assessing mediation in multilevel data: the advantages of multilevel SEM. *Struct. Equ. Model.* 18:161–82

Preacher KJ, Zhang Z, Zyphur MJ. 2016. Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychol. Methods* 21:189–205

Preacher KJ, Zyphur MJ, Zhang Z. 2010. A general multilevel SEM framework for assessing multilevel mediation. *Psychol. Methods* 15:209–33

Provides a framework and syntax to test for multilevel mediation within MSEM.

Purvanova RK, Bono JE, Dzieweczynski J. 2006. Transformational leadership, job characteristics, and organizational citizenship performance. *Hum. Perform.* 19:1–22

Quinn RE, Spreitzer GM. 1991. The psychometrics of the competing values culture instrument and an analysis of the impact of organizational culture on quality of life. *Res. Org. Change Dev.* 5:115–42

Rabe-Hesketh S, Skrondal A, Pickles A. 2004. Generalized multilevel structural equation modeling. *Psychometrika* 69:167–90

Raudenbush SW. 1997. Statistical analysis and optimal design for cluster randomized trials. *Psychol. Methods* 2:173–85

Rousseau DM. 1985. Issues of level in organizational research: multi-level and cross-level perspectives. In *Research in Organizational Behavior*, Vol. 7, ed. LL Cummings, B Staw, pp. 1–38. Greenwich, CT: JAI

Ryu E. 2011. Effects of skewness and kurtosis on normal-theory based maximum likelihood test statistic in multilevel structural equation modeling. *Behav. Res. Methods* 43:1066–74

Ryu E. 2014. Model fit evaluation in multilevel structural equation models. *Front. Psychol.* 5(81):1–9

Reviews level-specific tests and indices to assess fit in MSEM.

Ryu E, West SG. 2009. Level-specific evaluation of model fit in multilevel structural equation modeling. *Struct. Equ. Model.* 16:583–601

Scheipl F, Greven S, Küchenhoff H. 2008. Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Comput. Stat. Data An.* 52:3283–99

Scherbaum CA, Ferreter JM. 2009. Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organ. Res. Methods* 12:347–67

Schermelleh-Engel K, Kerwer M, Klein AG. 2014. Evaluation of model fit in nonlinear multilevel structural equation modeling. *Front. Psychol.* 5(181):1–11

Schneider B. 1987. The people make the place. *Per. Psychol.* 14:437–53

Selya AS, Rose JS, Dierker LC, Hedeker D, Mermelstein RJ. 2012. A practical guide to calculating Cohen's f^2, a measure of local effect size, from PROC MIXED. *Front. Psychol.* 3(111):1–6

Snijders TA., Bosker RJ. 1993. Standard errors and sample sizes for two-level research. *J. Educ. Behav. Stat.* 18:237–59

Snijders TA, Bosker RJ. 1994. Modeled variance in two-level models. *Socio. Meth. Res.* 22:342–63

Snijders TA, Bosker RR. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage

Sobel ME. 1982. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociol. Methodol.* 13:290–312

Stegmueller D. 2013. How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *Am. J. Polit. Sci.* 57:748–61

Teerenstra S, Moerbeek M, van Achterberg T, Pelzer BJ, Borm JF. 2008. Sample size calculations for 3-level cluster randomized trials. *Clin. Trials* 5:486–95

Tofighi D, Thoemmes F. 2014. Single-level and multilevel mediation analysis. *J. Early Adolesc.* 34:93–119

Tonidandel S, Williams EB, LeBreton JM. 2015. Size matters... Just not in the way that you think. See Lance & Vandenberg 2015, pp. 162–83

van der Leeden R, Busing FM, Meijer E. 1997. *Applications of bootstrap methods for two-level models*. Presented at Int. Multilevel Conf., 1st, Amsterdam, The Netherlands

Yuan KH, Bentler PM. 2007. Multilevel covariance structure analysis by fitting multiple single-level models. *Sociol. Methodol.* 37:53–82

Yuan Y, MacKinnon DP. 2009. Bayesian mediation analysis. *Psychol. Methods* 14:301–22

Yuan Y, MacKinnon DP. 2014. Robust mediation analysis based on median regression. *Psychol. Methods* 19:1–20

Zhang Z, Zyphur MJ, Preacher KJ. 2009. Testing multilevel mediation using hierarchical linear models problems and solutions. *Organ. Res. Methods* 12:695–719

Describes different mediation models, their potential confounding of between and within effects, and solutions.

Annual Review of Organizational Psychology and Organizational Behavior

Volume 4, 2017

Contents

Errata

An online log of corrections to *Annual Review of Organizational Psychology and Organizational Behavior* articles may be found at http://www.annualreviews.org/errata/orgpsych