Conducting and evaluating multilevel studies: recommendations, resources, and a checklist

Vicente González-Romá & Ana Hernández

Idocal, University of Valencia, Spain

Postprint

Reference:

González-Romá, V. & Hernández, A. (2023). Conducting and evaluating multilevel studies: recommendations, resources, and a checklist. *Organizational Research Methods, 26*(4), 629–654. DOI: 10.1177/10944281211060712

Correspondence concerning this article should be addressed to Vicente González-Romá, Idocal, Facultat de Psicologia, Universitat de València, Av. Blasco Ibáñez, 21, 46010-Valencia, Spain. E-mail: vicente.glez-roma@uv.es

Abstract

Multilevel methods allow researchers to investigate relationships that expand across levels (e.g., individual, team, organization). The popularity of these methods for studying organizational phenomena has increased in recent decades. Methodologists have examined how these methods work under different conditions, providing an empirical base for making sound decisions when using these methods. In this article, we provide recommendations, tools, resources, and a checklist that can be useful for scholars involved in conducting or assessing multilevel studies. The focus of our article is on two-level designs in which Level-1 entities are neatly nested within Level-2 entities, and top-down effects are estimated. However, some of our recommendations are also applicable to more complex multilevel designs.

Keywords: multilevel analysis, multilevel Structural Equation Modeling, cross-level effects, cross-level interaction, multilevel mediation, multilevel moderated mediation

Many organizational phenomena are multilevel because they involve variables that reside at different levels of analysis. To investigate relationships that span across levels, multilevel (ML)[1] modeling methods are needed. Thus, researchers interested in ML phenomena need to know how to deal with several key aspects involved in an ML study. Moreover, considering the increasing use of ML methods in our field (González-Romá & Hernández, 2017), reviewers need to be prepared to evaluate manuscripts that implement these methods. This requires understanding certain important issues and knowing some appropriate ways to handle them. Fortunately, research on ML methods is ripe enough to offer a set of recommendations (summarized in Table 1), several tools and resources (see Table 2), and an evaluation checklist (see Table 3), which can be useful to researchers who plan to conduct a multilevel study and reviewers and journal editors who frequently evaluate ML studies. Thus, the goal of the present article is to provide a set of recommendations and resources. We hope to contribute to the field by a) offering a comprehensive approach that covers the initial stages to the final stages of ML studies; b) helping researchers to make sound decisions when planning ML studies; c) increasing the rigor of ML studies; and d) facilitating reviewers' work when evaluating ML manuscripts. Due to space limitations, we focus on two-level designs in which Level-1 (L1) entities are neatly nested within Level-2 (L2) entities (e.g., employees nested within teams; departments nested within firms), and top-down effects are estimated. We do this because these are the most frequently used designs in our field (Molina-Azorín et al., 2020).

When and why we use multilevel methods

[1] Readers can see the list of abbreviations used in this article in the appendix.

Typically, researchers use ML modeling methods when the relationships investigated involve variables that reside at different levels. In these cases, researchers collect data about the study variables in a sample of L1 entities (e.g., individuals, departments) who belong to the sampled L2 units[2] (e.g., teams, firms, respectively). This results in a database with a nested structure.

Due to several factors (e.g., social interaction), employees in the same unit tend to have similar work experiences. Thus, nested data tend to show some degree of non-independence. Analyzing nested data by means of ordinary least squares (OLS) regression at the lower level can have undesirable consequences because the OLS assumption of independence of observations is violated (Heck & Thomas, 2015). In this regard, Bliese and Hanges (2004) showed that: i. estimating the relationship between an L2 variable and an L1 variable by using OLS regression leads to an increase in Type I error; and ii. estimating the relationship between two L1 variables using OLS regression and nested data leads to an increase in Type II error and a loss of statistical power (Bliese & Hanges, 2004). Furthermore, Bliese et al. (2018) showed that even a very low degree of non-independence (as indicated by an Intraclass Correlation Coefficient (ICC) = .013) affects the standard errors of parameter estimates. Thus, we recommend that researchers use ML modeling methods when analyzing data with a nested structure (Bliese et al., 2018).

Construct meaning

Generally, multilevel studies involve constructs specified at higher levels. It is extremely important to clarify the meaning of these constructs before formulating the study hypotheses and conducting the analyses (Chen et al., 2004; Jak, 2019; Preacher et al.,

[2] We use the term "unit" to refer to the different types of work-units that can be identified in organizations (e.g., team, department, organization).

2010). Without this clarification, it is not possible to fully and precisely interpret the empirical results obtained for these constructs and draw the subsequent conclusions.

Unfortunately, current practices in published studies do not reflect the importance of construct clarification. Kim et al.'s (2016) review concluded that "explicit discussions of how researchers conceptualize the constructs in their studies … at each level are lacking" (p. 892). ML researchers should take the construct meaning issue seriously. Hence, we propose that researchers address the following points:

1. *Provide an explicit definition of all the study constructs*, especially those residing at higher levels (Chen et al., 2005).

2. *Specify the nature of higher-level constructs*. Higher-level constructs can be of different types. A useful typology was proposed by Kozlowski and Klein (2000), who distinguished among: a) global unit properties, which are properties of the unit as a whole (e.g., unit size); b) shared unit properties, which describe characteristics that are common to unit members and originate in lower-level properties (e.g., team climate); and c) configural unit properties, which also originate in lower-level properties, but convey the pattern of individuals' experiences and attributes within a unit (e.g., climate uniformity).

3. *When necessary, explain how higher-level constructs emerge*. Some higher-level constructs originate in individuals' properties (e.g., perceptions, affect, behaviors). The latter combine through certain processes (e.g., social interaction) to yield higher-level constructs that have some features (e.g., sharedness, synergy, complementarity) that are not present in the corresponding individual elements (Eckardt et al., 2021). In these cases, it is necessary to explain how higher-level constructs emerge from individual properties to fully

understand the nature and foundation of the former[3]. Unfortunately, this explanation is frequently missing in research manuscripts (Eckardt et al., 2021; González-Romá, 2019). This explanation requires: 1. specifying the type of emergence involved, and 2. explaining the processes and factors involved in the emergence of higher-level constructs.

Kozlowski and Klein (2000) proposed an emergence typology with two general types, composition and compilation. Composition processes of emergence explain how convergence and within-unit agreement develop to yield a shared unit property. One of the psychosocial processes that explain convergence and within-unit agreement is social interaction (Ashforth, 1985). Compilation processes promote variability and configuration, and they explain how different types or/and amounts of individual-level properties combine to yield higher-level configural properties. One factor that may explain variability and configuration within units is demographic diversity (González-Romá & Hernández, 2014). Explaining how higher-level constructs emerge helps to understand the relationship between higher-level constructs and their individual-level counterparts. This relationship can also be clarified by using Chan's (1998) composition models.

4. *When ML models include isomorphic constructs, test for isomorphism*. ML isomorphism means that: i. "higher-level constructs have similar meanings and properties as their lower-level counterparts" (Tay et al., 2014, p. 78); and ii. both types of constructs show similar relationships with other constructs within an ML nomological network (Kozlowski & Klein, 2000). Generally, isomorphic constructs appear in homologies (i.e., ML models positing parallel relationships between constructs across levels). An often overlooked point

[3] Some higher-level constructs operationalized via aggregation do not require a theory of emergence because they do not imply new features (e.g., sharedness) emerging from the combination of individual properties (e.g., the aggregation of individual sales to operationalize team performance) (Eckardt et al., 2021). The emergence requirement will depend on the nature of the involved construct.

is that ML isomorphism requires psychometric isomorphism or measurement equivalence across levels (Jak, 2019; Tay et al., 2014). Psychometric isomorphism is crucial when higher-level constructs are formed following composition models of direct-consensus and referent-shift consensus (Chan, 1998). However, it is not required for additive, dispersion, or process composition models (see Tay et al., 2014, p. 85). Psychometric isomorphism involves ascertaining whether: i. the same dimensions underlie the investigated construct at different levels; and ii. factor loadings are invariant across levels. This isomorphism can be tested by ML factor analysis (see Tay et al., 2014). Note that if different dimensions underlie the studied construct at different levels, the dimensions used to describe the involved entities at different levels cannot be the same. If the factor loadings change across levels, the defining characteristics of the studied construct change across levels, and the construct cannot have the same interpretation across levels. Finally, we recommend taking the validity of constructs across levels seriously and implementing some of the different approaches proposed in the literature (see Chen et al., 2004; Tay et al., 2014).

Formulating multilevel hypotheses

Hypotheses specify the expected relationships between variables (Bacharach, 1989). When formulating hypotheses, researchers have to be aware of: i. the precise meaning of the variables involved in the statistical analysis conducted for hypothesis testing; and ii. what this analysis really does. This will ensure that the hypothesized relationships are aligned with the estimated relationships. This is especially important when formulating ML hypotheses because the variables and relationships mentioned in the hypotheses often do not completely match the variables and relationships modeled in the statistical analysis. In fact, current practice shows that we (researchers) frequently fail to formulate multilevel hypotheses that are fully aligned with the estimated relationships (see LoPilato &

Vandenberg, 2015; Bliese et al., 2018). To avoid this, a deeper understanding of what ML modeling methods really do in four specific cases can be helpful. We focus on these cases because they are quite common in ML research and offer room for improvement.

1. Hypotheses involving an individual-level predictor centered within cluster. Centering is a common practice that helps to interpret variable values by setting a reference zero point. When an L1 (e.g. individual) predictor's influence is of interest and a cross-level interaction effect is examined, the general recommendation is to center L1 predictors (X) around the group mean[4] (Aguinis et al., 2013; Enders & Tofighi, 2007; this practice is called centering within cluster (CWC) or group-mean centering). In these cases, centered values indicate subjects' standings on X relative to the unit mean, rather than an absolute value. CWC changes the meaning of values in L1 predictors. The associated ML hypotheses should acknowledge this change (Bliese et al., 2018). Thus, instead of hypothesizing that "At L1, X is positively/negatively related to Y", we should hypothesize that "At L1, subjects' relative X is positively/negatively related to subjects' relative Y".

2. Hypotheses about cross-level direct effects. The intercept-as-outcome ML model is popular among researchers. It is used to estimate cross-level direct effects (relationships between an L2 predictor and an L1 outcome). This model can be represented as follows:

L1 equation: $Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}$ (1)

L2 equations: $\beta_{0j} = \gamma_{00} + \gamma_{01} P_j + U_{0j}$ (2) $\beta_{1j} = \gamma_{10} + U_{1j}$ (3)

Y_{ij} is the score on the outcome of subject i from unit j, X_{ij} is the score on an L1 predictor of subject i from unit j, P_j is the score on an L2 predictor for each unit, β_{0j} and β_{1j} are the

[4] Cross-level interactions can also be investigated with grand-mean centering provided that the mean of the involved L1 predictor is reintroduced at L2 and its effect on the outcome and the interaction effect between this mean and the L2 moderator are controlled for (see Aguinis et al., 2013; Bliese et al., 2018).

regression intercept and slope, respectively, estimated in each unit (*j*), γ_{00} and γ_{10} are regression intercepts, γ_{01} is a regression slope, and r_{ij}, U_{0j}, and U_{0j} are residual terms.

Frequently, γ_{01} is interpreted as estimating the relationship between an L2 predictor (P_j) and the L1 outcome (Y_{ij}). However, this interpretation is not accurate (Bliese et al., 2018; LoPilato & Vandenberg, 2015). As equation (2) shows, γ_{01} estimates the relationship between an L2 predictor (P_j) and an L2 outcome (β_{0j}). Thus, to interpret γ_{01} accurately, the meaning of β_{0j} must be clarified. In this model, β_{0j} is a unit mean in the outcome (Y_{ij}), adjusted after controlling the effect of the unit mean in the predictor. Specifically, $\beta_{0j} = \overline{Y}_j - \beta_{1j}\ \overline{X}_j$[5] (see González-Romá, 2019; LoPilato & Vandenberg, 2015). Therefore, when hypothesizing cross-level direct effects, instead of hypothesizing that "P_j is related to Y_{ij}", we should hypothesize that "P_j is related to the units' mean in Y_{ij}".

3. Mediation hypotheses involving a higher-level variable. In nested data, the variance of variables measured at L1 can be decomposed into two orthogonal components: a between-cluster component and a within-cluster component[6] (Preacher et al., 2010). Variables measured at L2 (e.g., unit size) only have between components of variance. "Because Between and Within components are uncorrelated, it is not possible for a Between component to affect a Within component or vice versa" (Preacher et al., 2010, p. 210). Therefore, "any mediation effect in a model in which at least one of X, M, or Y [i.e., the predictor, the mediator, or the outcome] is assessed at Level 2 must occur strictly at the between-group level" (Preacher et al., 2010, p. 210). Thus, when researchers formulate

[5] The specific meaning of β_{0j} depends on the specific form of the multilevel model. As this equation shows, when there is no L1 predictor or when an included L1 predictor is group-mean centered, β_{0j} equals the (unadjusted) unit mean in the outcome.

[6] The proportion of variance of Y corresponding to these components can be estimated by computing ICC.

mediation hypotheses that involve an L2 variable, the hypothesized relationships among the between components of the involved variables should be specified.

4. Moderation hypotheses. Because L1 variables have between- and within-cluster components, when they appear in interaction terms it is extremely important to specify the component involved in the interaction. Depending on this component, the meaning of the interaction term and the corresponding moderation hypothesis may change (see Preacher et al., 2016). Fortunately, being aware of all the possible moderation effects in an ML design offers opportunities for theoretical development because it helps to uncover "hidden" moderations. Thus, we suggest that researchers think carefully about all the possible moderation effects existing in a given ML design, specify the within and between components involved, and focus on the ones dictated by their theoretical framework.

Deciding on Conventional ML modeling or ML Structural Equation Modeling

Although Conventional ML modeling (CMLM) and ML Structural Equation Modeling (MLSEM) are valid routes in ML research, the latter has several advantages. First, MLSEM can simultaneously account for measurement and sampling error (Marsh et al., 2009), whereas CMLM ignores both types of errors, which can bias the parameter estimates (Lütke et al., 2008, 2011). Second, MLSEM provides goodness-of-fit indices for each level of analysis (Ryu, 2014), whereas judging fit in CMLM is troublesome (Hox, 2010). Finally, MLSEM partitions the variance of L1 predictors into two orthogonal (between and within) latent components (Asparouhov & Muthén, 2019), whereas in CMLM the effects operating at different levels are conflated (e.g. Preacher et al., 2011; Zhang et al., 2009). The two sources of variance can be deconflated by CWC the L1 predictors and reintroducing the cluster means at L2 (a procedure known as CWC(M); Zhang et al., 2009). However, this latter approach still assumes that the observed means are perfectly reliable indicators of the L2 scores.

Despite the advantages of MLSEM, we do not suggest that MLSEM should replace CMLM. In fact, MLSEM has a major drawback: due to its complexity, it only performs well with larger samples. MLSEM shows more convergence problems (Li & Beretvas, 2013, Ludtke et al., 2011) and requires larger samples to reach similar power levels as CMLM (McNeish, 2017a; Zigler & Ye, 2019). In fact, small samples can often be more simply and effectively analyzed with CMLM (McNeish, 2017a). In addition, the choice may also depend on the types of variables modeled (Chen et al., 2004). For example, correcting for sampling error is an issue of concern when L1 variables are aggregated to operationalize L2 constructs (e.g., unit climate), but not for global L2 variables that have no L1 analogue (e.g., firm size). Similarly, measurement error is of particular concern when modeling constructs operationalized with several items responded to by individuals (e.g., unit culture), but it may be less important for variables such as salary or sales. Finally, neither CMLM nor MLSEM adequately deals with measurement error in dispersion constructs.

Thus, the choice between MLSEM and CMLM depends on sample size, model complexity, and the types of effects researchers want to test. MLSEM can generally be recommended if samples are large enough (i.e., a minimum of 100 L2 units with 15 subjects per unit –González-Romá & Hernández, 2017) or measurement and/or sampling error is an issue. For small samples, CMLM is recommended (McNeish, 2017a). However, if the model is too complex to be tested with CMLM, Bayesian MLSEM is recommended (Hox et al., 2012; Asparohouv & Muthén, 2019), especially with informative priors (e.g., Holtmann et al., 2016; McNeish, 2017a).

Data preparation and sample size

Before testing the study hypotheses, researchers need to consider several important issues: mean centering predictors, outliers, missing data, and sample size.

1. Mean-centering. When centering L1 predictors (including mediators and covariates), it is advisable to disentangle the between- and within-cluster components (Zhang et al., 2009). As mentioned earlier, in CMLM this is typically accomplished with CWC(M) (Enders & Tofighi, 2007; Zhang et al., 2009), which allows researchers to test and quantify the effects at both levels of analysis (Enders, 2013; LaHuis et al, 2019). If the interest is in directly estimating contextual effects (whether the relationship between the predictor and the outcome differs across levels), L1 predictors should be grand-mean centered[7], and their cluster means introduced at L2 (GMC(M)). The L2 slope captures the contextual effect, and the L1 slope represents the unconflated within effect (Enders, 2013; Hoffman, 2019). Regardless of the centering option, modeling the cluster means at L2 prevents bias due to omitted L2 variables (Antonakis et al., 2021; Bell et al., 2019). It is important to point out that the fact that cross-level and between-level (and contextual) effects can be analyzed by mean centering the L1 predictors and reintroducing the cluster means at L2 does not imply that an L2 construct exists (although this may be the case). L2 constructs that are operationalized from L1 data require a composition model to justify how higher-level constructs emerge and specify how the lower-level data should be combined to compose the higher-level construct (Kozlowski & Klein, 2000, van Mierlo et al., 2008).

When using MLSEM, the between and within variance components of L1 predictors are disentangled by latent mean centering (Asparouhov & Muthén, 2006a, 2019; Lüdtke et al., 2011). A simpler hybrid option is sometimes used for complex models, where only the

[7] In this case, raw or uncentered scores can also be used if there is a meaningful zero

between variance is modeled as a latent component (to correct for sampling error), while the L1 predictor is kept uncentered (Asparouhov & Muthén, 2019). Because centering occurs behind the scenes in MLSEM, researchers need to be aware that the default options may change depending on the estimation methods, software, and ML models (Asparouhov & Muthén, 2019, 2020; Hoffman, 2019). Thus, we strongly advise researchers to find out what these options are, in order to interpret the effects correctly.

2. Outliers. They can occur at different levels and bias ML results (Kloke et al., 2009; Pinheiro et al., 2001). Thus, outliers must be identified to assess whether they are errors to be corrected (e.g., sampling or coding errors) or meaningful outliers that influence ML results (Aguinis et al., 2013a; Langford & Lewis, 1998). In the latter case, researchers can delete outliers or use robust methods to reduce their impact (e.g., bootstrapping, heavy-tailed, or rank-based methods) (e.g., Aguinis et al., 2013a; Finch, 2017), but this impact should be assessed and explained (Aguinis et al., 2013a; Loy & Hoffman, 2013).

3. Missing data. Missing data models should be consistent with the specific ML statistical models tested; the former should include the effects considered in the latter (Grund et al., 2016; 2019; van Buuren, 2018). Consistency is achieved by employing estimation methods that use all the available data when fitting a model, such as Full Information Maximum Likelihood (FIML) (see Grund et al., 2019)[8], Fully Bayesian methods (Asparouhov & Muthén, 2019, 2020), or multiple imputation (MI). ML extensions of traditional MI work well for random intercepts and contextual effects (see Mistler & Enders, 2017). However, for random slopes, Fully Bayesian MI is recommended (Enders et al., 2020; Goldstain et

[8] FIML deals with missing data in outcome variables (for which distributional assumptions are assumed). This limitation can be solved by using MLSEM and defining predictors as endogenous variables (see Grund et al., 2019).

al., 2014). These methods are available in MI packages such as BLIMP (Keller & Enders, 2019) or JOMO (Quartagno et al., 2019).

4. Sample size recommendations. Deciding on the best combination of L1 and L2 sample sizes is a complex issue because it depends on many factors, such as the level of dependency in the data (ICC), the effect size, the estimation method, or the type of effect, among others. In general, simulations suggest that it is better to have more groups of fewer individuals than the other way around, for both CMLM and MLSEM. However, the latter is more demanding in terms of sample size. The reader can consult several reviews on sample size guidelines for different conditions and types of effects (González-Romá & Hernández, 2017; McNeish & Stapleton, 2016a; Hox & McNeish, 2020). These reviews show that CMLM typically offers unbiased and precise parameter estimates with samples as small as 20-30 L2 units of 5-10 cases each. However, it is more demanding in terms of power, especially for cross-level interactions. For example, Arend and Shäfer (2019) showed that, for medium ICCs, effect sizes, and slope variance components, adequate power levels (≥ .80) were reached with L2/L1 sample sizes of 40/3 or 30/5 (for L1 effects), and combinations ranging from 150/3 to 90/25 and from 200/9 to 125/25 (for cross-level direct effects and interactions, respectively). For MLSEM, the reviews mentioned above suggest that although 50 groups may suffice for small models, a minimum of 100 L2 units of 15-20 L1 units each is typically required to reach convergence and accurate estimates. If samples are smaller, Bayesian estimation is recommended (Asparohouv & Muthén, 2020; Zitzmann et al., 2016) with carefully selected priors (Depaoli & Clifton, 2015).

Although sample size guidelines are useful, they are based on specific conditions that may not generalize to the researcher's case. Thus, it is advisable to carry out power analysis to establish the sample sizes required at different levels (Scherbaum & Pesner, 2019)[9]. Although software based on approximate formulas can be used for simple models with fixed effects, Monte-Carlo-based simulation is the recommended strategy (e.g., Arend & Shäfer, 2019; Lane & Hennes, 2018; Sagan, 2019). In a priori analysis, different scenarios with different effects and sample sizes can be simulated to make a more informed decision about recommended sample sizes to reach enough power (see Arend & Shäfer (2019) for examples, guidelines, and recommendations). However, we acknowledge that a priori power analyses may be very hard to run with complex models.

Fitting an ML model

ML models are typically estimated using Maximum Likelihood methods (Hox et al., 2018)[10]. Particularly, in CMLM, FIML and REML (Restricted Maximum Likelihood) can be used, which are robust against mild violations of assumptions (e.g., non-normal residuals) when samples are large. With large samples, FIML is preferable to REML because it allows nested models that differ in fixed and/or random parts to be compared by means of chi-square tests (Hox, 1998). However, if the number of L2 units is small (i.e., less than 50 plus the number of L2 predictors –Snijders & Bosker, 2012), REML is recommended because it shows less bias in variance components (Hox et al., 2018; Hox & McNeish, 2020). Results of REML improve further if the Kenward-Roger correction is applied (McNeish, 2017a, 2017b).

In MLSEM, the conventional method is FIML (Hox et al., 2018). FIML is often combined with robust chi-squares and standard errors (Robust Maximum Likelihood - RML) if distributional assumptions are unmet (Hox et al., 2010). In fact, when normality is

[9] For an overview of Methods for Power Estimation in Two-Level Models, see Arend and Shäfer (2019)

[10] For brief and accessible introductions to different ML estimation methods and sample size requirements, see Hox et al. (2018) and McNeish and Stapleton (2016b)

seriously violated, robust standard errors are more precise, provided that samples are large (100 groups) (Maas & Hox, 2004). However, Hox et al. (2010) warned against the practice of using RML with small samples without testing distributional assumptions. When assumptions hold and data are continuous, RML only performs well with a large number of clusters (i.e., 200). This can be generalized to ordinal data with five or more categories (which are often assumed to be continuous and analyzed by RML; see Padget & Morgan, 2020). With fewer categories, other robust methods such as Diagonally Weighted Least Squares are preferred (Asparahouhov & Muthén, 2007; DiStefano & Morgan, 2014, Heck & Thomas, 2015).

When samples are small, models are intractable with maximum likelihood (e.g., random slopes and categorical items), or they show convergence issues, Bayesian estimation is recommended[11], both for CMLM and MLSEM. However, although Bayesian methods improve convergence rates (Depaoli & Clifton, 2015), the use of uninformative priors does not generally overcome Maximum Likelihood estimates in terms of bias and power (NcNeish, 2016), and it may even make them worse (McNeish, 2017a). Thus, informative priors should be chosen carefully (Bolin et al., 2019). However, informative priors do not have to be strong to be useful (McNeish, 2016). Weak priors are even preferred if it is unclear how to form strong ones (Depaoli & Clifton, 2015)[12].

One advantage of using MLSEM is that SEM programs provide a variety of indices to assess model fit. However, well-known fit indices designed for the single-level case

[11] For a primer on Bayesian estimation, see Jebb and Woo (2015) and Kaplan and Depaoli (2013). For a recent systematic review and comparison of different frequentist and Bayesian estimation methods in ML research with small samples, see Smid et al. (2020) and Zittman et al. (2020). For suggestions on prior construction, see Gelman (2006), McNeish and Stapleton (2016b), Smid et al. (2020), and Zittman et al. (2020).

[12] For suggestions on prior construction, see Gelman (2006), McNeish and Stapleton (2016b), Smid et al. (2020), and Zittman et al. (2020).

present two important problems in ML models: 1. Model fit assessment is dominated by model fit at the lower level because the sample size at this level is much larger; and 2. when the indices indicate a poor fit, it is not possible to determine the level where the reason for the model misfit resides. This situation led methodologists to derive procedures to obtain level-specific indices of model fit (e.g., Yuan & Bentler, 2007; Ryu & West, 2009). Some of them have been implemented in software packages (e.g., Mplus, Muthén & Muthén, 2017; OpenMx, Rappaport et al., 2020). We strongly recommend that researchers compute the available level-specific indices to assess the fit of MLSEM models.

Testing ML effects

Before testing ML effects such as cross-level direct effects and interactions, it is common to test whether there is enough variability across intercepts and slopes, respectively (Gavin & Hofmann 2002). When testing variability, the one-tail likelihood ratio test (see Hox et al., 2018) and the confidence intervals created around the variance estimated by Residual Bootstrap or Bayesian methods (see Aguinis et al, 2013b) are recommended. However, their results should not keep researchers from testing cross-level hypotheses (Aguinis et al. 2013b, LaHuis & Ferguson, 2009) due to low statistical power (Berkhof & Snijders 2001, LaHuis & Ferguson, 2009). Instead, ICC(1) and ICC(β) (Aguinis & Culpepper, 2015) can help to quantify the amount of variance attributed to intercept and slope differences, respectively.

Fixed effects are typically tested by means of the Wald test[13]. When cross-level interactions are significant, Preacher et al.'s (2006) tools are helpful for analyzing and interpreting the conditional effects. When the interest is in ML mediation, different types of

[13] Some authors argue that the Likelihood Ratio Test (LRT) strategy is a better option because the Wald test is more sensitive to model parameterization (see Hox et al., 2018).

indirect effects of a predictor X on an outcome Y via a mediator M are possible (depending on whether the variables reside at L1 or L2) (Bauer et al., 2006; Zhang et al., 2009). Regardless of the mediation model, indirect effects (which involve products of coefficients) do not distribute normally. The Monte Carlo-based Confidence Interval method is typically recommended to test for significance of the indirect effect (Fang et al., 2019; Tofighi & MacKinnon, 2011). Bayesian estimation (especially with informative priors) is also promising when samples are small (Fang et al, 2019). These recommendations also apply to ML conditional mediation models when conditional indirect effects are tested across different levels of the moderator (see Hayes & Rockwood, 2020). Table 2 shows a number of useful tools for these additional tests and plots for both CMLM and MLSEM.

Reporting ML analysis

To foster transparency and replicability, authors should provide information about their methodological decisions and justify their soundness. The recommendations provided in this paper should be considered. Moreover, when reporting ML results, researchers should strive to provide confidence intervals (Tonidandel et al., 2015), effect sizes [see Hammaker & Muthén (2020), LaHuis et al., (2019) and Rights & Sterba (2019)], and power levels (Scherbaum & Pesner, 2019). For more recommendations on reporting ML research, see Ferron et al. (2008), Jackson (2010), Monsalves et al. (2020), and Luo et al. (2021).

Conclusion

A limitation of this article is that we focused on a typical two-level design and did not consider other alternatives (e.g., designs with three levels, cross-classification of L1 entities, and bottom-up effects; see Heck et al., 2013; Preacher et al. 2010). However, because the two-level designs considered are quite popular in our field, we think the recommendations, tools, and resources presented will help to improve the quality of ML studies and facilitate reviewers' and editors' work.

References

Aguinis, H., & Culpepper, S. A. (2015). An expanded decision-making procedure for examining cross-level interaction effects with multilevel modeling. *Organizational Research Methods, 18*, 155-176.

Aguinis, H., Gottfredson, R. K., & Joo, H. (2013a). Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods, 16*, 270-301.

Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013b). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management, 39*, 1490-1528.

Antonakis, J., Bastardoz, N., & Rönkkö, M. (2021). On ignoring the random effects assumption in multilevel models: Review, critique, and recommendations. *Organizational Research Methods, 24*, 443-483

Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods, 24,* 1-19.

Ashforth, B. E. (1985). Climate formation: Issues and extension. *Academy of Management Review, 4*, 837–847.

Asparouhov, T., & Muthén, B. (2003). *Full-information maximum-likelihood estimation of general two-level latent variable models with missing data: A technical report*. Los Angeles: Muthén & Muthén.

Asparouhov, T., & Muthén, B. (2006, August). Multilevel modeling of complex survey data. *Proceedings of the Joint Statistical Meeting,* Seattle, WA. ASA section on Survey Research Methods, 2718–2726.

Asparouhov, T., & Muthén, B. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. *Joint statistical meeting: Section on statistics in epidemiology*, Salt Lake City, Utah

Asparouhov, T., & Muthén, B. O. (2010). *Multiple imputation with Mplus* (Technical Appendix). http://statmodel.com/download/ Imputations7.pdf

Asparouhov, T., & Muthén, B. (2019). Latent variable centering of predictors and mediators in multilevel and time-series models. *Structural Equation Modeling: A Multidisciplinary Journal, 26*, 119-142.

Asparouhov, T. & Muthén, B. (2020): Bayesian estimation of single and multilevel models with latent variable interactions, *Structural Equation Modeling: A Multidisciplinary Journal*. https://doi.org/10.1080/10705511.2020.1761808

Bacharach, S. B. (1989). Organizational theories: Some criteria for evaluation. *Academy of Management Review, 14*, 496-515.

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods, 11*, 142–163.

Bell, A., Fairbrother, M., & Jones, K. (2019). Fixed and random effects models: making an informed choice. *Quality & Quantity, 53*, 1051-1074.

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods, 18*, 151-164.

Biemann, T., Cole, M. S., & Voelpel, S. (2012). Within-group agreement: On the use (and misuse) of r_{wg} and $r_{wg(j)}$ in leadership research and some best practice guidelines. *The Leadership Quarterly, 23*, 66-80.

Bliese, P. D. & Hanges, P. J. (2004). Being both too liberal and too conservative: the perils of treating grouped data as though they were independent. *Organizational Research Methods, 7*, 400–17

Bliese, P. D., Maltarich, M. A., & Hendricks, J. L. (2018). Back to basics with mixed-effects models: Nine take-away points. *Journal of Business and Psychology, 33*, 1-23.

Bolin, J. H., Finch, W. H., & Stenger, R. (2019). Estimation of random coefficient multilevel models in the context of small numbers of level 2 clusters. *Educational and Psychological Measurement, 79*, 217-248.

Bosker, R. J., Snijders, T. A. B., & Guldemond, H. (2007). PinT (Power in two-level designs): Estimating standard errors of regression coefficients in hierarchical linear models for power calculations (Version 2.12). https://www.stats.ox.ac.uk/~snijders/multilevel.htm#progPINT

Browne, W.J.; Charlton, C.M.J.; Parker, R.M.A. (2019). *Developing a statistical analysis assistant using the Stat-JR software system version 1.0.7.* Centre for Multilevel Modelling, University of Bristol, UK.

Browne, W. J., Lahi, M. G., & Parker, R. M. (2009). A *guide to sample size calculations for random effect models via simulation and the MLPowSim software package*. University of Bristol. http://www.bristol.ac.uk/cmm/software/mlpowsim/

Chan, D. (1998). Functional relationships among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology, 83*, 234–246.

Chen, G., Mathieu, J., & Bliese, P. (2004). A framework for conducting multi-level construct validation. *Multi-level Issues in Organizational Behavior and Processes, Research in Multi-Level Issues, 3*, 273-303.

Chen, G., Smith, T. A., Kirkman, B. L., Zhang, P., Lemoine, G. J., & Farh, J. L. (2019). Multiple team membership and empowerment spillover effects: Can empowerment processes cross team boundaries?. *Journal of Applied Psychology, 104,* 321-340

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal, 22*, 327-351.

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling: A Multidisciplinary Journal, 21*, 425-438.

Eckardt, R., Yammarino, F. J., Dionne, S. D., & Spain, S. M. (2021). Multilevel methods and statistics: The next frontier. *Organizational Research Methods, 24*, 187-218.

Enders, C. K. (2013). Centering predictors and contextual effects. In M. A. Scott, J. S. Simonoff, & B. D. Marx (Eds.), *The Sage handbook of multilevel modeling* (pp. 89–109). Sage.

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

Enders, C.K., Keller, B.T., & Levy, R. (2018). A fully conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods, 23*, 298-317

Enders, C. K., Du, H., & Keller, B. T. (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological Methods, 25*, 88–112.

Fang, J., Wen, Z., & Hau, K. T. (2019). Mediation effects in 2-1-1 multilevel model: evaluation of alternative estimation methods. *Structural Equation Modeling: A Multidisciplinary Journal, 26*, 591-606.

Ferron, J. M., Hogarty, K. Y., Dedrick, R., Hess, M., Niles, J., & Kromrey, J. D. (2008). Reporting results from multilevel analysis. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data.* Information Age.

Finch, H. (2017). Multilevel modeling in the presence of outliers: A comparison of robust estimation methods*. Psicologica: International Journal of Methodology and Experimental Psychology, 38*, 57-92.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological methods, 19*, 72.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis, 1*, 515-534.

Goldstain, H. (2014). REALCOM-IMPUTE: multiple imputation using MLwin. http://www.bristol.ac.uk/media-library/sites/cmm/migrated/documents/imputation.pdf

Goldstain, H ., Carpenter, J. R., & Browne, W. J. (2014). Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. *Journal of the Royal Statistical Society: Series A, 177*, 553-564.

González-Romá, V. (2019). Three issues in multilevel research. *The Spanish Journal of Psychology, 22*, e4, 1–7.

González-Romá, V., & Hernández, A. (2014). Climate Uniformity: Its Influence on Team Communication Quality, Task Conflict, and Team Performance. *Journal of Applied Psychology, 99*, 1042–1058.

González-Romá, V., & Hernández, A. (2017). Multilevel modeling: Research-based lessons for substantive researchers. *Annual Review of Organizational Psychology and Organizational Behavior, 4,* 183-210.

González-Romá, V., Peiró, J. M., & Tordera, N. (2002). An examination of the antecedents and moderator influences of climate strength. *Journal of Applied Psychology, 87*, 465–473.

Green, P., & Macleod, C. J. (2016a). *Package "SIMR".* Retrieved from https://cran.r-project.org/web/packages/simr/index.html

Green, P., & Macleod, C. J. (2016b). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7,* 493-498.

Grund, S., Lüdtke, O., & Robitzsch, A. (2016). Multiple imputation of missing covariate values in multilevel models with random slopes: A cautionary note. *Behavior Research Methods, 48,* 640-649.

Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods, 21*, 111-149.

Grund, S., Lüdtke, O., & Robitzsch, A. (2019). Missing data in multilevel research. (2019). Explained variance measures for multilevel models. In S. E. Humphrey & J. M. LeBreton (Eds.), *The handbook of multilevel theory, measurement, and analysis* (p. 353–364). American Psychological Association

Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods, 25*, 365.

Hayes, T. (2019). Flexible, free Software for multilevel multiple imputation: A review of Blimp and jomo. *Journal of Educational and Behavioral Statistics, 44*, 625-641.

Hayes, A. F., & Rockwood, N. J. (2020). Conditional process analysis: Concepts, computation, and advances in the modeling of the contingencies of mechanisms. *American Behavioral Scientist, 64*, 19-54.

Heck, R. H. & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques: MLM and SEM Approaches Using Mplus*. Routledge.

Heck, R. H., Thomas, S. L., & Tabata, L. N. (2013). *Multilevel and longitudinal modeling with IBM SPSS*. Routledge.

Hoffman, L. (2019). On the interpretation of parameters in multivariate multilevel models across different combinations of model specification and estimation*. Advances in Methods and Practices in Psychological Science, 2*, 288-311.

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

Holtmann, J., Koch, T., Lochner, K., & Eid, M. (2016). A comparison of ML, WLSMV, and Bayesian methods for multilevel structural equation models in small samples: A simulation study. *Multivariate behavioral research, 51*, 661-680.

Hox, J. J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147–154). Springer Verlag.

Hox, J.J. (2010). *Multilevel Analysis: Techniques and Applications.* Routledge. 2nd ed.

Hox, J. J., & Maas, C. J. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural equation modeling: A Multidisciplinary Journal, 8*, 157-174.

Hox, J., & McNeish, D. (2020). Small samples in multilevel modeling. In R.van de Schoot & M.Milocevic (Eds.), *Small sample size solutions* (pp. 215–225). Routledge.

Hox, J. J., Maas, C. J., & Brinkhuis, M. J. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica neerlandica, 64*, 157-170.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications (3^rd^).* Routledge.

Hox, J. J., Van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods, 6*, 87–93.

Jak, S. (2019). Cross-level invariance in multilevel factor models. *Structural Equation Modeling: A Multidisciplinary Journal, 26*, 607-622.

Jackson, D. L. (2010). Reporting results of latent growth modeling and multilevel modeling analyses: Some recommendations for rehabilitation psychology. *Rehabilitation Psychology, 55*, 272-285.

Jebb, A. T., & Woo, S. E. (2015). A Bayesian primer for the organizational sciences: The "two sources" and an introduction to BugsXLA. *Organizational Research Methods, 18*, 92-132.

Kaplan, D., & Depaoli, S. (2013). Bayesian statistical methods. In T. D. Little (Ed.), *Oxford handbook of quantitative methods* (pp. 407–437). Oxford University Press.

Keller, B. T., & Enders, C. K. (2019). *Blimp User's Manual (Version 2.1).* http://www.appliedmissingdata.com/blimpusermanual-2-1.pdf

Kim, E. S., Dedrick, R. F., Cao, C., & Ferron, J. M. (2016). Multilevel factor analysis: Reporting guidelines and a review of reporting practices. *Multivariate Behavioral Research, 51*, 881–898

Kloke, J. D., McKean, J. W., & Rashid, M. M. (2009). Rank-based estimation and associated inferences for linear models with cluster correlated errors. *Journal of the American Statistical Association, 104*, 384-390.

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

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

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate behavioral research*, *36*, 249-277.

LaHuis, D. M., & Ferguson, M. W. (2009). The accuracy of significance tests for slope variance components in multilevel random coefficient models. *Organizational Research Methods, 12*, 418-435.

LaHuis, D. M., Blackmore, C. E., & Bryant-Lees, K. B. (2019). Explained variance measures for multilevel models. In S. E. Humphrey & J. M. LeBreton (Eds.), *The handbook of multilevel theory, measurement, and analysis* (p. 353–364). American Psychological Association.

Lai, M. H. C. (2019). *bootmlm: Bootstrap resampling for multilevel models* (R Package Version 0.0.1doi:10.5281/zenodo.1879127

Lai, M. H. (2020). Bootstrap confidence intervals for multilevel standardized effect size. *Multivariate Behavioral Research, 1-21.*

Lane, S. P., & Hennes, E. P. (2018). Power struggles: Estimating sample size for multilevel relationships research. *Journal of Social and Personal Relationships, 35*, 7-31.

Langford, I. H., & Lewis, T. (1998). Outliers in multilevel data. *Journal of the Royal Statistical Society: Series A*, *161*, 121-160

Li, X., & Beretvas, S. N. (2013). Sample size limits for estimating upper level mediation models using multilevel SEM. *Structural Equation Modeling: A Multidisciplinary Journal, 20*, 241-264.

LoPilato A. C., & Vandenberg R. J. (2015). The not-so-direct cross-level direct effect. In C. E. Lance &R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* (pp. 292–310). Routledge.

Lorah, J. (2018). Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. *Large-Scale Assessments in Education, 6,* 8.

Loy, A., & Hofmann, H. (2013). Diagnostic tools for hierarchical linear models. *Wiley Interdisciplinary Reviews: Computational Statistics, 5*, 48-61.

Lüdtke, O., Marsh, H. W., 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. *Psychological Methods, 16*, 444–467.

Lüdtke, O., Marsh, H. W., 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. *Psychological methods*, *13*, 203-229.

Luo, W., Li, H., Baek, E., Chen, S., Lam, K. H., & Semma, B. (2021). Reporting Practice in Multilevel Modeling: A Revisit After 10 Years. *Review of Educational Research, 91,* 311–355

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*, 86-92.

Mai, Y., & Zhang, Z. (2018). Software packages for Bayesian multilevel modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 25*, 650-658.

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research, 44*, 764-802.

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

McCoach, D. B., Rifenbark, G. G., Newton, S. D., Li, X., Kooken, J., Yomtov, D., & Bellara, A. (2018). Does the package matter? A comparison of five common multilevel modeling software packages. *Journal of Educational and Behavioral Statistics, 43*, 594-627.

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal,* 23, 750–773.

McNeish, D. (2017a) Multilevel mediation with small samples: a cautionary note on the multilevel structural equation modeling framework. *Structural Equation Modeling: A Multidisciplinary Journal, 24*, 609-625.

McNeish, D. (2017b). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral Research, 52*, 661-670.

McNeish, D. M., & Stapleton, L. M. (2016a). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review, 28*, 295-314.

McNeish, D., & Stapleton, L. M. (2016b). Modeling clustered data with very few *clusters. Multivariate behavioral research, 51*, 495-518.

Mistler, S. A., & Enders, C. K. (2017). A comparison of joint model and fully conditional specification imputation for multilevel missing data. *Journal of Educational and Behavioral Statistics, 42*, 432-466.

Molina-Azorín, J. F., Pereira-Moliner, J., López-Gamero, M. D., Pertusa-Ortega, E. M., & José Tarí, J. (2020). Multilevel research: Foundations and opportunities in management. *Business Research Quarterly, 23*, 319-333.

Monsalves, M. J., Bangdiwala, A. S., Thabane, A., & Bangdiwala, S. I. (2020). LEVEL (Logical Explanations & Visualizations of Estimates in Linear mixed models): recommendations for reporting multilevel data and analyses. *BMC medical research methodology, 20*.

Muthén, B. O., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. Hox & J. K. Roberts (Eds.), *The handbook of advanced multilevel analysis* (pp. 15–40). Taylor & Francis

Muthén, L.K. & Muthén, B.O. (2017). *Mplus User's Guide. Eighth Edition*. Muthén & Muthén

Padgett, R. N., & Morgan, G. B. (2020). Multilevel CFA with Ordered Categorical Data: A Simulation Study Comparing Fit Indices Across Robust Estimation Methods. *Structural Equation Modeling: A Multidisciplinary Journal.* https://doi.org/10.1080/10705511.2020.1759426

Pinheiro, J. C., Liu, C., & Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics, 10*, 249-276.

Pituch, K. A., & Stapleton, L. M. (2012). Distinguishing between cross-and cluster-level mediation processes in the cluster randomized trial. *Sociological Methods & Research, 41*, 630-670.

Preacher, K. J., & Selig, J. P. (2010). *Monte Carlo method for assessing multilevel Mediation: An interactive tool for creating confidence intervals for indirect effects in 1-1-1 multilevel models* [Computer software]. http://quantpsy.org/.

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of educational and behavioral statistics*, *31*, 437-448.

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel SEM. *Structural Equation Modeling: A Multidisciplinary Journal, 18*, 161-182.

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

Preacher, K. J., Zhang, Z. & Zyphur, M. J. (2016). Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychological Methods, 21*, 189–205.

Quartagno, M., Grund, S., & Carpenter, J. (2019). Jomo: a flexible package for two-level joint modelling multiple imputation. *R Journal.* https://discovery.ucl.ac.uk/id/eprint/10078316/1/RJwrapper.pdf

Rappaport, L. M., Amstadter, A. B., & Neale, M. C. (2020). Model Fit Estimation for Multilevel Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal, 27*, 318-329.

Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X., Martinez, A., Bloom, H., Hill, C. (2011). Optimal design software for multi-level and longitudinal research (Version 3.01) [Software] https://www.wtgrantfoundation.org

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological methods, 24*, 309.

Rights, J. D., Preacher, K. J., & Cole, D. A. (2019). The danger of conflating level-specific effects of control variables when primary interest lies in level-2 effects. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12194

Ryu, E. (2014). Model fit evaluation in multilevel structural equation models. *Frontiers in psychology, 5,* 81.

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*, 583-601.

Sagan, A. (2019). Sample size in multilevel structural equation modeling–the Monte Carlo approach. *Econometrics, 23*, 63-79.

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

Scherbaum, C. A., & Pesner, E. (2019). Power analysis for multilevel research. In S. E. Humphrey & J. M. LeBreton (Eds*.), The handbook of multilevel theory, measurement, and analysis* (p. 329–352). American Psychological Association.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin, 86*, 420-428.

Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal, 27*, 131-161.

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

Tay, L., Woo, S. E., & Vermunt, J. K. (2014). A conceptual and methodological framework for psychometric isomorphism: Validation of multilevel construct measures. *Organizational Research Methods, 17*, 77-106.

Tingley, D, Yamamoto, T, Hirose, K, Keele, L, Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software, 59*, 1-38

Tingley D,. Yamamoto T., Hirose K., Keele L., Imai, K., Trinh, M. & Wong, W. (2019). *Causal Mediation Analysis.* https://cran.r-project.org/web/packages/mediation/mediation.pdf

Tofighi, D., & MacKinnon, D. P. (2011). RMediation: An R package for mediation analysis confidence intervals. *Behavior Research Methods, 43*, 692-700.

Tonidandel, S., Williams, E. B., & LeBreton, J. M. (2014). Size matters... just not in the way that you think. In C. E. Lance & R. J.Vandenberg (Eds). *More statistical and methodological myths and urban legends* (pp. 162-183). Routledge.

Van Buuren, S. (2018). Flexible imputation of missing data. CRC press.

Van der Leeden, R., Meijer, E., & Busing, F. M. (2008). Resampling multilevel models. In J. Leeuw & Meijer , E. (Eds) *Handbook of multilevel analysis* (pp. 401-433). Springer.

Van Mierlo, H., Vermunt, J. K., & Rutte, C. G. (2009). Composing group-level constructs from individual-level survey data. *Organizational Research Methods, 12,* 368-392.

Vandenberg, R. J., Richardson, H. A. (2019). A primer on multilevel structural modeling: User-friendly guidelines. In Humphrey, S., LeBreton, J. (Eds.), *The handbook of multilevel theory, measurement, and analysis* (pp. 449–472). American Psychological Association.

Vuorre, M., (2017). bmlm: Bayesian Multilevel Mediation. R package version 1.3.4. https://cran.r-project.org/package=bmlm

Wolak, M. E., Fairbairn, D. J., & Paulsen, Y. R. (2012). Guidelines for Estimating Repeatability. *Methods in Ecology and Evolution, 3,* 129-137.

Yuan, K. H., & Bentler, P. M. (2007). 3. Multilevel Covariance Structure Analysis by Fitting Multiple Single-Level Models. *Sociological methodology, 37*, 53-82.

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. Psychological methods, 14, 301.

Zhang, Z., Zyphur, M. J., & Preacher, K. J. (2009). Testing multilevel mediation using hierarchical linear models: Problems and solutions. *Organizational Research Methods, 12*, 695-719.

Zigler, C. K., & Ye, F. (2019). A Comparison of Multilevel Mediation Modeling Methods: Recommendations for Applied Researchers. *Multivariate behavioral research, 54*, 338-359.

Zitzmann, S., Lüdtke, O., Robitzsch, A., & Marsh, H. W. (2016). A Bayesian approach for estimating multilevel latent contextual models. *Structural Equation Modeling: A Multidisciplinary Journal, 23*, 661-679.

Zitzmann, S., Lüdtke, O., Robitzsch, A., & Hecht, M. (2020). On the Performance of Bayesian Approaches in Small Samples: A Comment on Smid, McNeish, Miocevic, and van de Schoot (2020). *Structural Equation Modeling: A Multidisciplinary Journal.* https://doi.org/10.1080/10705511.2020.1752216

Table 1

Multilevel topics and their corresponding recommendations

Topics	*Recommendations*	*References*
When and why ML methods are used	• When the data analyzed have a nested structure (no matter whether the relationships investigated span across level or do not) ML methods allow researchers to deal with nonindependence of data	Bliese & Hanges (2004) Bliese et al. (2018)
Construct meaning and emergence	• Provide an explicit definition of higher-level constructs • Specify the nature of higher-level constructs • When needed, explain how higher-level constructs emerge: - specify the type of emergence involved - explain the psychosocial processes and factors involved in the emergence of higher-level constructs - explain the relationship between higher-level constructs and their individual-level counterparts • Test for psychometric isomorphism when needed	Chan (1998) Chen et al. (2004) González-Romá (2019) Jak (2019) Kozlowski & Klein (2000) Tay et al. (2014)
Elaborating multilevel hypotheses	• Adjust the formulation of ML hypotheses to: i. the precise meaning of variables, and ii. what ML analysis really does. • Pay attention to the following cases: - an individual-level predictor (X) is centered within cluster: note that centered values represent subjects' standings on X relative to the group mean, not the absolute value. - a hypothesis about a cross-level direct effect is formulated: note that the outcome variable is a(n adjusted) mean[a] in the outcome variable (Y), not the individual values in the outcome variable. - a mediation hypothesis involving a higher-level variable is formulated: the expected relationships should be specified among the between components of the involved variables	Bliese et al. (2018) González-Romá (2019) LoPilato & Vandenberg (2015) Preacher et al. (2010) Preacher et al. (2016)

	- a moderation hypothesis is formulated: think carefully about all the possible moderation effects, specify the within and between components involved, and focus on those dictated by the adopted theoretical framework	
Deciding on CMLM or MLSEM[b]	• MLSEM can be typically recommended if samples are large enough considering the model's complexity (i.e. a minimum of 100 L2 units with 15-20 subjects per unit). For smaller samples, or if sampling/measurement error is not a serious concern, researchers should a) Use the modified versions of CMLM by using CWC(M) or GMC(M)[c], depending on the research hypotheses to unconflate L1 slopes. Or, b) Use Bayesian MLSEM • Some additional issues must be considered when choosing between CMLM or MLSEM: - For 2-1-1 and 2-2-1 models, where the indirect effects are "between" effects, MLSEM is preferred if the number of clusters is large enough (at least 50 for 2-1-1 models and 80 for 2-2-1 models) - For 1-1-1 models with random slopes, the CWC(M) is recommended when either a very small within indirect effect is expected or a negative covariance between the random coefficients could suppress the within indirect effect.	González-Romá & Hernández (2017) Hox et al. (2012) Li & Beretvas (2013) McNiesh (2017a) Ziegler & Ye (2019)
Centering L1 predictors	• In CMLM, the within- and between-level effects of L1 predictors (including mediators and covariates) should be unconflated and differentiated by CWC(M) or GMC(M). The choice depends on: - Whether researchers want the L2 effects to capture between or contextual effects. In the former case, CWC(M) is the best option. In the latter, GMC(M) allows a direct test of contextual effects • When the interest is in cross-level interactions in which it is assumed that L1 slopes vary at random and depend on an L2 moderator, CWC can be used. However, in this case, the between-portion variance of L1 scores (which may interact with the L2 moderator) is ignored. Thus, the typical recommendation is to use CWC(M) – although GMC(M) is also a valid option.	Aguinis et al., (2013) Asparouhov & Muthén (2019) Enders & Tofighi (2007) Hoffman (2019) Hoffman & Gavin (1998) Rights et al. (2019) Zhang et al. (2009)

	• If Grand Mean Centering (GMC) instead of GMC(M) is used, the L1 effect will be a conflated mixture of within and between variances. GMC should be avoided unless researchers have sound reasons to do it. • In MLSEM, latent mean centering is typically recommended. For complex models (e.g. with random slopes or cross-level interactions), latent mean centering requires Bayesian estimation methods. If researchers want to use Maximum Likelihood methods in complex models, the hybrid centering option should be used. If the sampling ratio approaches 100%, there are no missing data, and cluster sizes are large, centering based on observed means (CWC(M)) should be used because it works better than using latent means	
Detecting and managing outliers	• Check whether there are outliers in the initial database and, when present, analyze the influence of these outliers by comparing the results with those obtained either by deleting outliers or by minimizing their impact by means of robust techniques. • Explain observed differences, if any	Aguinis et al (2013a) Finch (2017) Loy & Hoffman (2013)
Handling Missing Data	• To handle missing data, use estimation methods that employ all available information in the data (FIML or Full Bayesian), or use ML multiple imputation methods, which should be congenial to the research model (i.e. the imputation model should take into account the nested structure of the data and include all the parameters included in the statistical model to be tested). • Listwise deletion should be avoided, especially when missing data are observed in the predictors and covariates and the missing mechanism is not completely at random.	Asparouhov & Muthén (2020) Grund et al. (2019) Hayes (2019)
Sample sizes and power	• Check whether L1 and L2 sample sizes are large enough to test the hypotheses involved in the research model (e.g. cross-level moderation, mediation, etc.), according to existing simulations, and considering the ML approach (CMLM or MLSEM) and the estimation method (e.g. FIML, REML, Bayesian). “With smaller samples: keep the model simple” (Hox & McNeish, 2020; pp. 221-222) • Whenever possible (i.e. if the model complexity allows this), plan the minimum sample size required to have enough power to detect ML effects using existing software.	González-Romá & Hernández (2017) McNeish & Stapleton (2016a) McNeish (2017a) Hox & McNeish (2020). Hox et al., (2018) Lane & Hennes (2018) Mathieu et al. (2012) Arend & Shäffer (2019)

	• After carrying out the analyses, estimate and report the actual power levels attained by means of existing software or Monte-Carlo simulations	
Fitting an ML model	• Choose the most adequate estimation method considering sample size, distributional assumptions, and model complexity. When using frequentist methods: - In CMLM use FIML estimation methods if samples are large enough (i.e. at least 50 L2 units plus the number of L2 predictors; Snijders & Bosker, 2012). For smaller samples, use REML and a Kenward-Roger correction if possible. - In MLSEM, use FIML if distributional assumptions hold and samples are large (typically 100 L2 units of 15-20 individuals). If distributional assumptions are seriously violated and items are continuous or approach continuity, use Robust standard errors and chi-square tests. For categorical items, use methods based on Weighted Least Squares. • If samples are small, models are intractable with maximum likelihood, or they do not converge in proper solutions, use Bayesian Estimation methods, and whenever possible, use informative priors. • The estimation methods available (and the default methods), as well as the particular corrections to obtain robust standard errors, depend on the particular software used. Check the reference manuals (and updates) for the particular version of the software to be used • When using MLSEM, assess model fit at each level.	Asparahouhov & Muthén (2007) Depaoli & Clifton (2015) Hox et al. (2010) Hox & McNeish (2020) NcNeish, (2017a, 2017b) Ryu (2014) Ryu & West (2009) Yuan & Bentler (2007)
Testing effects	• Quantify the proportion of criterion variance attributed to intercept and slope differences by means of ICC(1) and ICC(β). Despite the power problems for detecting random effects, if researchers want to test whether this variability is statistically significant, Wald's test should be avoided. The one-tail likelihood ratio test, Residual Bootstrap, or Bayesian methods are better alternatives. • When testing cross-level moderation effects, plot and test for conditional effects and regions of significance. • When testing for mediation and moderated mediation, use adequate tests that do not assume that the indirect effects (and the conditional indirect effects) follow a	Aguinis & Culpepper (2015) Aguinis et al. (2013b) Fang et al. (2019) González-Romá & Hernández (2019) Hox et al. (2018)

	normal distribution, such as Monte-Carlo based confidence intervals or Bayesian estimation. For moderated mediation, test for conditional indirect effects.	
Reporting	• Provide detailed information about the methodological decisions made and justify their soundness, and consider the following issues: 1. Construct operationalization: a) measurement instruments and adaptations; b) aggregation procedures for construct operationalization (e.g. ICCs and emergence); and c) psychometric quality (reliability and validity, and when necessary, measurement equivalence) aligned with the levels of analysis. 2. Outlier detection and management 3. Missing data treatment 4. Centering methods used 5. Model specification, estimation methods, and software. If Bayesian methods are used, provide details of the prior distributions and the methods used to select them 6. Apart from the statistical significance of the parameter estimates, provide confidence intervals, effect sizes, power estimates, and, when possible, goodness-of-fit at each level	Bladwin & Fellingham, (2013). Ferron et al., (2008) Geldhof et al. (2014) Jak (2019) Jackson (2010) LaHuis et al. (2019) Monsalves et al. (2020)

Note. [a] The specific interpretation of the associated intercept depends on the specific model being tested and the centering procedure used.

[b] For a primer on MLSEM with Mplus syntax and examples, see Vandenberg and Richardson (2019)

[c] GMC(M): Grand-Mean Centering with cluster means introduced as L2 predictors

[d] Contextual=Between-Within. Thus, regardless of the centering option, both between and contextual effects can be obtained and tested.

Table 2

Multilevel tools and resources

Objective	
To compute ICCs:	• **ICC(1):** Rpackage "ICC" (Wolak et al., 2012) https://cran.r-project.org/web/packages/ICC Excel tool referenced in Biemann et al. (2012) • **ICC(β):** Rpackage "ICCbeta" (Aguinis & Culpepper, 2015) https://cran.r-project.org/package=iccbeta
To impute ML missing data	• Package 'micemd' (Audigier et al., 2018) https://www.rdocumentation.org/packages/micemd/versions/1.6.0 https://stefvanbuuren.name/fimd/sec-level2pred.html • REALCOM-Impute (Goldstein, 2014) http://www.bristol.ac.uk/cmm/software/realcom/imputation.html • BLIMP (Enders et al. 2018, 2020; Keller & Enders, 2019) http://www.appliedmissingdata.com/multilevel-imputation.html • JOMO (Quartagno et al., 2019) https://cran.r-project.org/web/packages/jomo • Stat-JR (Browne et al., 2019) http://www.bristol.ac.uk/cmm/research/missing-data/ • Mplus (Muthén & Muthén, 2017) TYPE = IMPUTATION command (Asparohouv & Muthén, 2010) • For recommendations depending on the types of effects to be tested and examples using different software packages, see Table 6 of Grund et al. (2018) • For recent reviews on ML multiple imputation, see Grund et al. (2019) and van Buuren (2018)
To run power analysis and determine sample	• Optimal Design (Raudenbush et al., 2011) http://hlmsoft.net/od/

size requirements to reach acceptable power	• PinT (Bosker et al., 2007) https://www.stats.ox.ac.uk/~snijders/multilevel.htm#progPINT • MLPowSim (Browne et al. 2009). http://www.bristol.ac.uk/cmm/software/mlpowsim/ • ML-power (Mathieu et al., 2012) https://aguinis.shinyapps.io/ml_power/ • R package SIMR (Green, & Macleod (2016a, 2016b; see also Arend & Shäffer, 2019). https://cran.r-project.org/web/packages/simr/index.html • For Mplus syntax examples to conduct a Monte Carlo simulation to estimate power, see Lane and Hennes (2018) • For a recent review on power analyses and sample size in multilevel models, see Scherbaum and Pesner (2019).
To estimate effect sizes	• r2mlm: R-Squared Measures for Multilevel Models (Rights & Sterba, 2019) https://CRAN.R-project.org/package=r2mlm • R package bootmlm: Bootstrap Confidence Intervals for ML Standardized Effect Size (Lai, 2019; 2020) https://rdrr.io/github/marklhc/bootmlm/man/bootmlm.html
To fit a ML model and assess goodness-of-fit	• For a comparison of different common programs that can fit ML models, see McCoach et al. (2018). • For a detailed review of the capabilities and characteristics of the programs that support Bayesian ML analyses, see Mai and Zhang (2018) • For recommendations about how to build priors when using Bayesian estimation, see Gelman (2006), Smid et al. (2020), and Zittman et al. (2020) • For computing fit indices at different levels, use Yuan & Bentler (2007) syntax (http://www3.nd.edu/~kyuan/multilevel/Multi-Single.sas) or programs such as Mplus (Muthén & Muthén, 2017) and OpenMx (Rappaport et al., 2020)
To test and plot ML moderation effects:	• Interactive calculation tools for establishing simple intercepts, simple slopes, and regions of significance (Preacher et al., 2006) http://www.quantpsy.org/interact/hlm2.htm

	• Interplot https://cran.r-project.org/web/packages/interplot/vignettes/interplot-vignette.html • Mplus syntax using the LOOP option and PLOT option in the MODEL CONSTRAINT command • Supplemental materials by Preacher et al. (2016) for MLSEM with Mplus http://quantpsy.org/pubs/preacher_zhang_zyphur_2016_(code.appendix).pdf
To test for indirect effects and conditional indirect effects (ML moderated mediation)	• MLmed macro in SPSS (Rockwood & Hayes, 2020): https://njrockwood.com/mlmed • Supplemental materials by Bauer et al. (2006) for SAS, SPSS and HLM: http://dx.doi.org/10.1037/1082-989X.11.2.142.supp http://www.quantpsy.org/pubs/bpg_2006_supp_spss.zip http://www.quantpsy.org/pubs/bpg_2006_supp_hlm.zip • RMediation package (Tofighi & MacKinnon, 2011) https://CRAN.R-project.org/package=RMediation https://amplab.shinyapps.io/MEDMC/ • Preacher & Selig's (2010) calculator http://quantpsy.org/medmc/medmc111.htm • Causal Mediation analysis (Tingley et al., 2014: 2019) https://CRAN.R-project.org/package=mediation • Supplemental materials by Zyphur et al (2019) for MLSEM with Mplus http://quantpsy.org/pubs/zyphur_zhang_preacher_bird_supp.zip
For Bayesian Multilevel Mediation	• Vourre (2017) https://cran.r-project.org/package=bmlm

Table 3

Checklist for evaluating multilevel studies

Do the authors …	Yes	No	Not applicable
1. Justification			
1.1. Explain why they (do not) use multilevel modeling methods?			
2. Construct meaning and emergence			
2.1. Provide explicit definitions of the study's higher-level constructs?			
2.2. Specify the nature of the investigated higher-level constructs?			
2.3. Explain, when needed, how the specified higher-level constructs emerge?			
- Specify the type of emergence involved?			
- Explain the psychosocial processes and factors involved in the emergence of higher-level constructs?			
- Explain the relationship between higher-level constructs and their individual-level counterparts?			
- Test for psychometric isomorphism when the research model includes isomorphic constructs?			
3. Elaborating multilevel hypotheses			
3.1. Adjust their ML hypotheses to: i. the precise meaning of variables, and ii. what ML analysis really does?			
- Correctly formulate hypotheses involving a L1 predictor (X) that has been centered within cluster, showing that the centered values represent subjects' standings on X relative to the unit mean?			
- Correctly formulate hypotheses about a "cross-level direct effect", showing that the outcome variable is an a(n adjusted) mean[a] in the outcome variable?			
- Correctly formulate mediation hypotheses involving a higher-level (L2) variable, showing that the expected relationships involve the between components of the studied variables?			
- Specify the moderation effects being tested by clarifying the within and between components of the predictor and moderator involved?			
4. Choosing between CMLM and MLSEM			

4.1. Justify their choice considering the research hypotheses (i.e. the types of effects to be tested and the types of constructs -aggregate or global- of interest)?

4.2. Justify their choice considering the recommendations about sample sizes at different levels and the effects of interest?

5. Centering L1 predictors[b]

5.1. If raw data or grand-mean centering is used, provide a sound justification for not disentangling within and between variance sources? (e.g. Chen et al., 2019)

5.2. Disentangle the between and within effects when using CMLM?

5.3. Justify their centering choice considering the study hypotheses?

5.4. Adequately interpret the parameter estimates to match the centering option used?

6. Managing Outliers

6.1. Assess whether there are meaningful outliers?

6.2. Indicate the method used to detect outliers?

6.3. When meaningful outliers are detected …

- Indicate how they were addressed?
- Compare the results with and without outliers' influence and provide an explanation for different results, if any?

7. Handling Missing data

7.1. Report the proportion of missing data at different levels?

7.2. Handle missing data by either

- Using estimation methods that utilize all available information and make it possible to handle the observed missing data (for a particular level, predictor, or outcome), or
- Imputing missing data using multiple imputation models that are congenial to the statistical ML model?
- If using multiple imputation, do authors report the software used for this purpose?

8. Considering the adequacy of Sample Sizes

8.1. Provide evidence that the sample size is reasonable according to existing simulation studies, considering:

- The analytical approach (CMLM or MLSEM)?
- The ML effects of interest (L1 effects, cross-level direct and interaction effects, mediation, etc.)?

- The estimation method (e.g. FIML, REML, Bayesian)?

8.2. Carry out power analysis before data collection (if the complexity of the model allows for it) to safeguard that the study sample is large enough to reach an acceptable power?

9. Fitting the ML model

9.1.Indicate the software used to test the research model?

9.2. Provide adequate justification for the estimation method used, considering:

- The sample size?
- The effects of interest?
- The satisfaction of distributional assumptions?

9.3. Describe the priors and the reasons to use these priors if Bayesian estimation methods are used?

9.4. Provide information about whether the model converged in a proper solution?

9.5. Explain how convergence/estimation problems, if any, were solved?

9.6. Assess model fit at each level when MLSEM is fitted?

10. Testing and quantifying the hypothesized ML effects

10.1. Clearly explain what variables are included in the fixed and random parts of the model, including control variables and interaction terms?

10.2. Indicate the particular tests used (e.g. Wald test, Likelihood Ratio Test, Monte Carlo) taking into account recommendations depending on the types of effects tested?

10.3. Provide Standard Errors and Confidence intervals for the parameters of interest?

10.4. Provide indicators of the size of the effects of interest?

10.5. Provide information about power?

10.6. Qualify the effects tested by considering the results of power analysis and effect sizes?[c]

10.7. When testing **moderation effects**…

- Focus on the right within and/or between components of the moderation depending on the level at which the predictors and the moderators are located.
- Provide additional information about the tested effect through a graphical representation that shows how it changes across the range of the moderator values with the corresponding significance region?

10.8. When testing **mediated or indirect effects**…

- Focus on the right within and/or between components of the indirect effects, depending on the level at which the predictors and the mediators are located?

- Estimate the right indirect effect, considering whether or not the paths involved in mediation vary at random within the L2 units?
- Test for significance of the indirect effects by means of methods that do not assume a normal distribution?

10.9 When testing **moderated mediation or conditional indirect effects**

- Focus on the right within and between components of the moderation depending on the levels at which the predictors, the mediators, and the moderators are located?
- Estimate the right conditional indirect effect, considering whether the paths involved in mediation vary at random?
- Test for significance of the conditional indirect effects by using methods that do not assume a normal distribution?
- Provide additional information about the conditional indirect effects through a graphical representation that shows how effects change across the range of the moderator values with the corresponding significance regions?

Note. Checklists are useful tools. However, they must be used with some flexibility because some items may not apply to some specific situations.

[a] The specific interpretation of the associated intercept depends on the specific model being tested and the centering procedure used.

[b] L2 predictors can only be centered by using GMC (this should be done if zero has not a meaningful interpretation)

[c] For example, a non-significant effect should be trusted more or less depending on whether the power is high enough or not (e.g. Mathieu et al., 2012), in combination with the effect size (LaHuis et al., 2019). If the effect is considered relevant in practice, and power is low, studies should cross-validate the results with larger samples. Some indirect ways of increasing power (e.g. adding relevant covariates, using more reliable measurement instruments) can also be used (Mathieu et al., 2012; Pituch & Stapleton 2012, Scherbaum & Ferreter 2009).

Appendix

List of abbreviations used in the article (in alphabetical order)

CMLM	Conventional Multilevel Modeling
CWC	Centering Within Cluster
CWC(M)	Centering Within Cluster with reintroduction of cluster means
FIML	Full Information Maximum Likelihood
GMC	Grand Mean Centering
GMC(M)	Grand Mean Centering with reintroduction of cluster means
ICC	Intraclass Correlation Coefficient
L1	Level-1
L2	Level-2
MI	Multiple Imputation
ML	Multilevel
MLSEM	Multilevel Structural Equation Modeling
OLS	Ordinary Least Squares
REML	Restricted Maximum Likelihood
RML	Robust Maximum Likelihood
SEM	Structural Equation Modeling