TPM®

MICRO-MACRO AND MACRO-MICRO EFFECT ESTIMATION IN SMALL SCALE LATENT VARIABLE MODELS WITH CROON'S METHOD

BEN KELCEY
FANGXING BAI
YANLI XIE
UNIVERSITY OF CINCINNATI

KYLE COX
UNIVERSITY OF NORTH CAROLINA CHARLOTTE

Psychological theories and research often incorporate and investigate how macro- and micro-level processes cooperate to convey, provision, and envelope effects. A historical challenge in these multilevel contexts has been incorporating micro-macro, bottom-up, or emergent effects alongside the more common top-down or macro-micro effects. Although multilevel structural equation modeling provides a framework for such analyses, a persistent issue is the large sample size requirements necessary to reliably estimate parameters. In this study, we outline extensions to the recently developed Croon-based estimator for multilevel structural equation models. We then evaluate the performance of Croon's approach under a method-of-moments corrected maximum likelihood estimator to probe models that integrate micro-macro and macro-micro effects. The results suggest that Croon's method often outperforms maximum likelihood in terms of convergence, bias, and root mean-squared error and represents a useful complementary estimator. We provide *R* code that applies the estimator to an example using the *lavaan* package.

Keywords: Multilevel structural equation modeling; Croon's method; Micro-macro effects; Organizational psychology; Leadership.

Correspondence concerning this article should be addressed to Ben Kelcey, University of Cincinnati, 2820 Bearcat Way, Cincinnati, Ohio, United States of America. Email: ben.kelcey@gmail.com

Across a broad swath of psychological research, there has been a growing interest in incorporating and investigating how macro- and micro-level processes cooperate to influence outcomes. For instance, organizational psychology has developed a range of multilevel theories on the performance of work teams that integrate a leader's ability to motivate team members (top-down or macro-micro effects) with how the quality of team member interaction patterns and shared mental models influences team outcomes (bottom-up or micro-macro effects; e.g., Kozlowski, Gully, Nason, & Smith, 1999; Marks, Sabella, Burke, & Zaccaro, 2002; Mohammed & Dumville, 2001). Recent research in this area has also identified theoretical, analytical, and empirical work that bridges these types of micro and macro domains to address top-down and bottom-up effects. This research has also identified these lines of inquiry as one of the most important challenges in management research (e.g., Aguinis, Boyd, Pierce, & Short, 2011; Mathieu & Chen, 2011). Similarly, in education, contemporary approaches to improve teacher and school effectiveness have emphasized the influence of individual teacher knowledge and skill on student achievement (top-down effects) while concurrently leveraging collaborative models of professional development to inculcate shared educational goals and a school-wide culture of continuous improvement (bottom-up effects; e.g., Desimone, 2009; Kelcey & Phelps, 2013).

More generally, theories across psychology often call upon complex, dynamic, and emergent processes to describe the interaction and formation of cognitive, behavioral, and emotional outcomes (e.g., Hilpert & Marchand, 2018). Prior theoretical and methodological research has detailed two related but distinct types of multilevel processes (e.g., Croon & van Veldhoven, 2007; Kozlowski & Klein, 2000; Lachowicz, Sterba, & Preacher, 2015). The first and most well-known is the top-down or macro-micro processes that describe how the characteristics of organizations or macro units influence the outcomes of individuals (e.g., Raudenbush & Bryk, 2002). The second set of processes is the bottom-up, emergent or micro-macro processes that describe how the interactions among individuals emerge to influence the outcomes of organizations (e.g., Croon & van Veldhoven, 2007).

Despite theoretical interest in integrating top-down and bottom-up multilevel processes, the preponderance of empirical and methodological research both within and outside of psychology has fixated on top-down or macro-micro effects (e.g., Cronin, Weingart, & Todorova, 2011). This top-down emphasis has been in part driven by the inability of common methods to appropriately incorporate bottom-up, emergent or micro-macro effects alongside the more common top-down or macro-micro effects (Croon & van Veldhoven, 2007). Developments in multilevel structural equation modeling have however reduced this disparity (e.g., Lüdtke, Marsh, Robitzsch, & Trautwein, 2011). The multilevel structural equation modeling framework provides a flexible approach for models that integrate top-down and bottom-up processes because it facilitates the decomposition of latent variables across hierarchical levels in ways that connect and quantify cross-level processes (Hox, Maas, & Brinkhuis, 2010; Lachowicz et al., 2015; Lüdtke et al., 2011).

Nevertheless, the practical utility of the multilevel structural equation modeling framework has often been tempered by model complexity and sample size considerations (e.g., Hox et al., 2010; Jak, 2019; Jak, Oort, & Dolan, 2014; Hox & McNeish, 2020). A key constraint in multilevel structural equation modeling is that estimation of parameters typically requires fairly large sample sizes to dependably deliver stable and admissible estimates of parameters (e.g., Hox et al., 2010; Jak, 2019; Li & Beretvas, 2013; Wolf, Harrington, Clark, & Miller, 2013). For instance, past methodological research has suggested that 100 or more clusters are needed to deliver stable and admissible estimates with relatively simple multilevel structural equation models (Hox & Maas, 2001; Hox et al., 2010; Kelcey, Cox, & Dong, 2019; Li & Beretvas, 2013; Smid & Rosseel, 2020). Investigations with even moderately complex multilevel structural equation models have suggested even larger samples may be required — often hundreds of clusters (Hox et al., 2010). In a broader context, the regularity with which estimation issues arise in multilevel structural equation models has become somewhat well-known and has led to interest in, development of, and use of alternative estimators that are more stable for samples of less than 100 clusters (e.g., Asparouhov & Muthen, 2007; Hox et al., 2010; Jak, 2019; Muthen, 1989, 1994; Smid & Rosseel, 2020; Takane, & Hwang, 2018; Yuan & Hayash, 2005).

The large sample size requirements of multilevel structural equation modeling has at times been particularly challenging for many areas of psychology where routine access to hundreds of groups and individuals is prohibitively expensive or unfeasible. For instance, in the area of educational psychology, samples of 100 or more schools and students are uncommon (e.g., Spybrook, Shi, & Kelcey, 2016). Moreover, across most areas inside and outside of psychology, there has been a growing mandate for empirical studies to establish evidence regarding the theories of action and effects of context. For instance, increasingly studies are charged with sourcing more comprehensive evidence that probes for whom and under what contexts effects manifest (e.g., Dong, Kelcey, & Spybrook, 2018; Moss, Kelcey, & Showers, 2014) while also detailing the intermediate processes and contexts that convey, provision, and envelope effects (e.g., Ilgen, Hollenbeck, Johnson, & Jundt, 2005; Kelcey, Dong, Spybrook, & Cox, 2017; Kelcey, Dong, Spybrook, & Shen, 2017).

For many studies, the sample sizes demanded by the complexity of macro- and micro-level processes implied by theories quickly overtakes the range of plausible sample sizes commonly found in many fields (e.g., Gagne & Hancock, 2006; Hox et al., 2010; Hox, Moerbeek, & van der Schoot, 2017). Furthermore, estimation with complex multilevel models can often be further complicated by the (im)balance of individual-level samples across groups (e.g., Guittet, Ravaud, & Giraudeau, 2006). Prior research has consistently demonstrated that parameter estimation can be particularly difficult when the number of individuals per team or cluster varies across clusters (e.g., Hox & Maas, 2001; Muthen, 1994; Raudenbush & Bryk, 2002). This research has found that the impact of unbalanced samples on estimation can be particularly large when the number of teams is small to moderate (e.g., Guittet et al., 2006). The net implication is that many small to moderate sample multilevel analyses encounter estimation issues and errors that force researchers to either test a simplified version of the guiding theory of action or rely on estimators or approaches that have with less desirable properties (Hox et al, 2010; Hox & McNeish, 2020; Kelcey, Cox, & Dong, 2019; Loncke, et al., 2018; Smid & Rosseel, 2020).

The ostensible way to circumvent such issues is to simply collect more data. In many areas of exploratory and experimental psychology, however, the development of theories to address multifarious lines of questions combined with calls for explanatory models has given rise to structural equation models whose complexity often outpaces sample size (e.g., Christ, Sibley, & Wagner, 2012; Hill, 2006; Hox et al., 2010). The imbalance between model complexity and available sample size can be particularly pronounced in multilevel settings where practical sampling constraints suggest that samples of 100 or more clusters (e.g., teams, corporations, schools) are often considered prohibitively large while statistical standards for multilevel structural equation modeling often suggest that samples of 100 clusters is only small to moderate in scale when it comes to parameter estimation (e.g., Hox et al, 2010; Schochet, 2011; Spybrook et al., 2016). Prior research has, however, asserted the foundational and theoretical value of investigating macro- and micro-level processes in small to moderate sized studies (e.g., Bodner & Bliese, 2017; Walton, 2014).

To address the disparity between model complexity and study scale that often arises in single and multilevel structural equation models, literature over the past several decades has developed a range of alternative estimators (e.g., Asparouhov & Muthen, 2007; Croon & van Veldhoven, 2007; Depaoli & Clifton, 2015; Hox et al, 2010; Jak, 2019; Muthen, 1989; 1994; Kelcey, Hill, & Chin, 2019; Takane & Hwang, 2018; Yuan & Hayash, 2005). Recent work along these lines has developed a new limited information estimator that introduces a bias-correction factor to overcome the shortcomings of many of the previous estimators (Croon & van Veldhoven, 2007; Devlieger, Mayer, & Rosseel, 2016; Devlieger & Rosseel, 2017; Kelcey, Cox, & Dong, 2019; Rosseel, 2020). This Croon-based estimator was designed to provide more stability in estimation while retaining minimal bias in small samples. The estimator draws on a form of bias-corrected factor score path analysis whose corrections were outlined in Croon (2002). Conceptually, the estimator exploits the simplicity and inherent stability of estimating a collection of smaller and simpler local measurement models, then amasses the results of the models, and finally corrects for the expected biases that arises from the piecewise approach.

There is a growing research base that has demonstrated the potential of Croon's estimation method across a variety of settings and model types (e.g., Devlieger et al., 2016; Devlieger & Rosseel, 2017; Hayes & Usami, 2019; Kelcey, 2019; Kelcey, Cox, & Dong, 2019; Loncke et al., 2018; Lu, Kwan, Thomas &

Cedzynski, 2011; Smid & Rosseel, 2020). For example, Croon-based estimation has demonstrated robust performances in standard single level structural equation models, structural equation models with cross-loadings in the measurement models, structural equation models with measurement and structural misspecifications, structural equation models with non-normal error distributions, and multilevel structural equation models (Devlieger et al., 2016; Devlieger & Rosseel, 2017; Hayes & Usami, 2019; Kelcey, 2019; Loncke et al., 2018; Lu et al., 2011).

In this study, we extended the scope of this work by investigating the corrections and performance of Croon's bias-corrected estimator in the context of multilevel structural equation models that integrate bottom-up and top-down effects. Below, we first develop a context and working example. We then describe the Croon-based estimation process and corrections for multilevel structural equation models that integrate bottom-up and top-down effects. We follow with an illustration of the method by applying it to our working example and provide *R* code to implement the analyses using the *lavaan* package (Rosseel, 2012). We then probe the performance of Croon's method relative to maximum likelihood estimation and (uncorrected) factor score path analysis using a Monte Carlo simulation. We end with a discussion.

CROON-BASED ESTIMATION

Working Example

We detail the Croon-based estimator within the context of an example study from the organizational psychology literature. Our example takes up an investigation regarding how team leaders' behaviors and their team members' perceptions of those behaviors support organizational change through an interplay of bottom-up and top-down effects (Nohe, Michaelis, Menges, Zhang, & Sonntag, 2013). A broad array of organizational studies has sought to identify and delineate the pivotal behaviors of effective leaders and the pathways through which these behaviors act on team performance outcomes (e.g., Walter & Bruch, 2009). One set of behaviors prominently highlighted in the literature is the charisma of leaders or the ability of leaders to project "symbolic leader influence rooted in emotional and ideological foundations" (Antonakis, Fenley, & Liechti, 2011, p. 376).

Prior research has consistently implicated charismatic behavior by leaders as an important attribute in cultivating organizational change and eventually higher team performance (e.g., DeGroot, Kiker, & Cross, 2000; Wu, Tsui, & Kinicki, 2010). However, much less is known regarding how leaders' charisma actually comes to improve team performance (Nohe et al., 2013). For these reasons, recent research has developed organizational change theories that interweave top-down and bottom-up processes to empirically delineate how effects are conveyed. One recent example of this line of inquiry is the theory presented and evaluated in Nohe et al. (2013). The guiding theory suggests that leaders' charisma shapes members' perceptions of their leadership (i.e., top-down effects) in ways that bolster team members' commitment to change (lateral effects) and ultimately improve team performance (bottom-up effects).

The theory of action that we examined in our illustrative example focuses on the system of relationships among the four core latent constructs depicted in Figure 1. This system employs latent variables that characterize both team- and individual-level constructs while drawing on a series of top-down and bottom-up processes. The first construct in our system focused on leader's change-promoting behaviors (*CPB*). A Leader's change-promoting behavior was the primary independent variable and was assessed at the team-level using the degree to which leaders engaged in six change-promoting behaviors (Herold, Fedor, Caldwell, & Liu, 2008). The second construct in this example system focused on individual team member's perceptions of leader charisma (*PLC*). Team member's perceptions of leader charisma served as the proximal or first mediator in the theory and the latent variable was measured using three questions regarding leader's idealized influence (e.g., Felfe, 2006). The third construct we draw on captured individual team member's commitment to change (*CTC*). Team member's commitment to change operated as the distal or second mediator in the system and was evaluated using four questions regarding the degree to which they were committed to change (e.g., Herscovitch & Meyer, 2002). The final construct in our example system assessed the team-level performance from leaders' perspectives (*TP*). Team-level performance from leaders' perspectives served as the targeted outcome and was evaluated using four questions (e.g., Conger, Kanungo, & Menon, 2000).

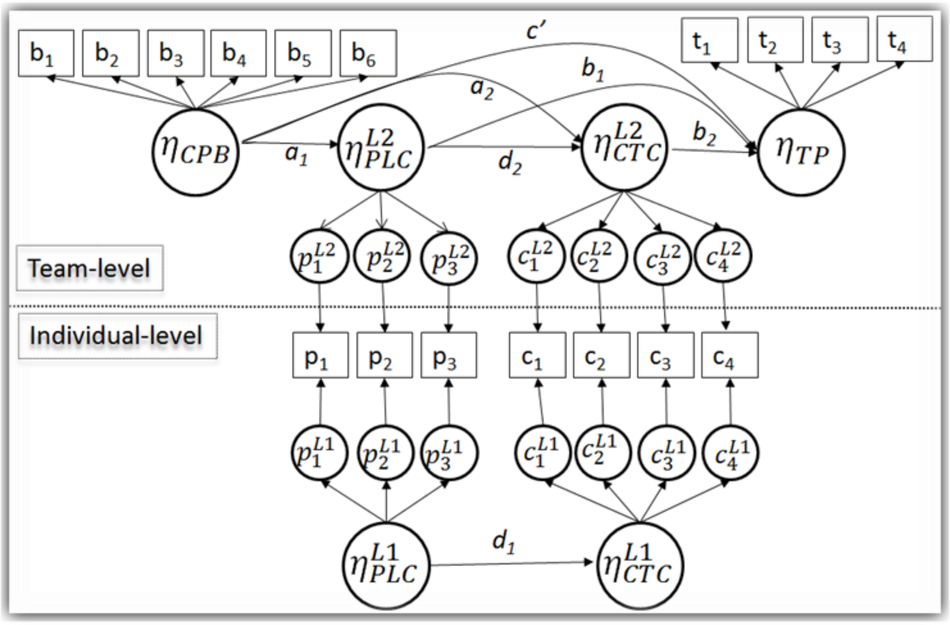

FIGURE 1
Multilevel diagram of the theory and structural paths of the working example connecting leader's change-promoting behaviors (*CPB*), individual team member's perceptions of leader charisma (*PLC*), individual team member's commitment to change (*CTC*), and team-level performance from leaders' perspectives (*TP*) together with the observed indicators for each construct.

Figure 1 provides a multilevel diagram of the theory and outlines the structural paths connecting the constructs together with the observed indicators for each construct. More abstractly, Figure 1 illustrates a case of sequential multilevel mediation that is characterized by team- and individual-level constructs that integrates top-down and bottom-up processes. The theory describes how a team-level variable (leader's change-promoting behaviors) influences an individual-level proximal mediator — team member's perceptions of leader charisma (top-down effect) — in ways that convey influence on an individual-level distal mediator — team member's commitment to change (lateral effect) — and ultimately manifest as improvements in a team-level outcome — team-level performance from leaders' perspectives (bottom-up effect).

Croon Estimation

We next delineate the theory of action described in Figure 1 as a multilevel structural equation model. Conceptually, a multilevel framework begins by decomposing the variation in the indicators used in Figure 1 (i.e., the collection of indicators labeled *b*, *p*, *c*, and *t* in Figure 1) into orthogonal components that represent variation across teams and variation across individuals within teams (e.g., Muthen, 1989; 1994). With $\mathbf{y}$ as the collection of indicators, the decomposition can be expressed as:

$$\begin{aligned} \mathbf{y} = \mathbf{y}^{L2} + \mathbf{y}^{L1} \quad & \mathbf{y}^{L2} \sim MVN(0, \boldsymbol{\Sigma}_{\mathbf{y}}^{L2}) \\ & \mathbf{y}^{L1} \sim MVN(0, \boldsymbol{\Sigma}_{\mathbf{y}}^{L1}) \end{aligned} \tag{1}$$

Here $\mathbf{y}^{L2}$ represents the team-level means of the indicators whereas $\mathbf{y}^{L1}$ captures the individual-specific deviations of an individual from his/her team mean. Under this decomposition, the covariances among the team-level indicator means is described using $\boldsymbol{\Sigma}_{\mathbf{y}}^{L2}$ whereas the covariance among the individual deviations from the indicator means within teams (individual-level) are described using $\boldsymbol{\Sigma}_{\mathbf{y}}^{L1}$.

For instance, applied to the team member's perceptions of leader charisma (*PLC*) construct in our working example, we can decompose the variation in each of the indicators such that $\mathbf{p}^{L2}$ captures the extent to which leaders differ along their team members' collective perceptions of their charisma whereas $\mathbf{p}^{L1}$ captures the differences between each individual's perception and the average perception of the leader across fellow team members. In turn, $\boldsymbol{\Sigma}_{\mathbf{p}}^{L2}$ describes how the average reported perceptions for each charisma indicator covary with the average reported perceptions of the other charisma indicators. Likewise, $\boldsymbol{\Sigma}_{\mathbf{p}}^{L1}$ describes how an individual's deviations in reported charisma for an indicator covary with that individual's deviations on other charisma indicators.

Once the variation in the indicators is split across levels, we can then develop measurement models that relate the indicators to their respective factors and structural models that relate the factors as specified in a substantive theory (e.g., see Figure 1). At the individual-level the measurement and structural models can be specified as:

$$\begin{aligned} \mathbf{y}^{L1} &= \boldsymbol{\Lambda}^{L1}\,\boldsymbol{\eta}^{L1} + \boldsymbol{\zeta}^{L1} \\ \boldsymbol{\eta}^{L1} &= \mathbf{B}^{L1}\,\boldsymbol{\eta}^{L1} + \boldsymbol{\varepsilon}^{L1} \end{aligned} \tag{2}$$

We use $\boldsymbol{\eta}^{L1}$ as the individual-level latent variables (e.g., η_{PLC}^{L1} and η_{CTC}^{L1} in Figure 1), $\boldsymbol{\Lambda}^{L1}$ as the individual-level factor loading patterns of the indicators and $\boldsymbol{\zeta}^{L1}$ as the individual-level residual errors of the indicators. For the structural models, we use $\mathbf{B}^{L1}$ as individual-level path coefficients connecting the latent variables and $\boldsymbol{\varepsilon}^{L1}$ as the individual-level regression residual errors. In Figure 1, for example, the $\mathbf{B}^{L1}$ set of path coefficients consists of d_1 that captures an individual-level relationship that connects individual team member's perceptions of leader charisma (*PLC*) with individual team member's commitment to change (*CTC*).

Similarly, at the team-level, the measurement and structural models can be specified as

$$\begin{aligned} \mathbf{y}^{L2} &= \boldsymbol{\Lambda}^{L2}\,\boldsymbol{\eta}^{L2} + \boldsymbol{\zeta}^{L2} \\ \boldsymbol{\eta}^{L2} &= \mathbf{B}^{L2}\,\boldsymbol{\eta}^{L2} + \boldsymbol{\varepsilon}^{L2} \end{aligned} \tag{3}$$

We use $\boldsymbol{\eta}^{L2}$ as the team-level latent variables (e.g., $\boldsymbol{\eta}_{\boldsymbol{CPB}}$, $\boldsymbol{\eta}_{\boldsymbol{PLC}}^{\boldsymbol{L2}}$, $\boldsymbol{\eta}_{\boldsymbol{CTC}}^{\boldsymbol{L2}}$, and $\boldsymbol{\eta}_{\boldsymbol{TP}}$ in Figure 1), $\boldsymbol{\Lambda}^{L2}$ as the team-level indicator factor loadings and $\boldsymbol{\zeta}^{L2}$ as the team-level residual errors of the indicators. Structural models then employ $\mathbf{B}^{L2}$ as the team-level path coefficients connecting the latent variables (e.g., coefficients a_1, a_2, b_1, b_2, c', and d_2 in Figure 1) and $\boldsymbol{\varepsilon}^{L2}$ as the residual errors.

Under this multilevel structural equation modeling framework, we can implement Croon-based estimation of the top-down and bottom-up parameters with (method-of-moments corrected) maximum likelihood using four conceptual steps. The first step is to individually estimate each of the common factor models implied by Figure 1. Our implementation and subsequent analyses focus on Croon's corrections under maximum likelihood. We note, however, Croon's approach and corrections are not specific to maximum likelihood and its method-of-moments based corrections can be adapted for other estimators (e.g., Devlieger & Rosseel, 2019; Rosseel, 2020). We draw on maximum likelihood estimation for single-level confirmatory factor analysis models for latent variables that are evaluated using team-level indicators (i.e., *CPB* and *TP*) and multilevel confirmatory factor analysis models for latent variables that draw on individual-level indicators (i.e., *PLC* and *CTC*).

The second step leverages the factor models to predict factor scores (e.g., using the regression method) and the resulting variances and covariances of the latent variables. Our implementation draws on the regression method for factor score prediction. However, the method can be readily implemented with alternative factor score prediction methods as well (e.g., Devlieger et al., 2016; Kelcey, Cox, & Dong, 2019). More generally, implementation of Croon's approach does not actually require factor score prediction (Devlieger & Rosseel, 2019). Croon's method can fully circumvent the use of factor scores by drawing on the implied covariances of indicators across latent variables. The use of factor scores is only a conceptually convenient and accessible way to describe the logic of the method.

When theories incorporate both top-down and bottom-up relationships, we need to predict factor scores at each level and form variance-covariance matrices at the team- and individual-level. For instance, in our application, the team-level variance-covariance matrix captures the team-level variance and covariance components among all four latent variables whereas the individual-level variance-covariance matrix captures the individual-level variance and covariance components between the *PLC* and *CTC* latent variables only.

In the third step, we correct the factor score variance-covariance matrices using the results of the factor models from step one. More specifically, the use of predicted factor scores to estimate the variance and covariances among latent variables neglects score uncertainty and results in biased estimates. As a result, we must leverage the parameter estimates stemming from the measurement models in step one (Equations 2 and 3) to infer the expected bias of the covariances introduced by predicting factor scores. In the context of multilevel models involving top-down and bottom-up relationships, the corrected covariance terms in the team-level covariance matrix ($\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{L2}$) can be estimated using:

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{L2} = (\mathbf{R}_{\boldsymbol{\eta}}^{L2})^{-1} \boldsymbol{\Sigma}_{f_{\boldsymbol{\eta}}}^{L2} (\mathbf{R}_{\boldsymbol{\eta}}^{L2})^{-T} \tag{4}$$

Here we use $\boldsymbol{\Sigma}_{f\boldsymbol{\eta}}^{L2}$ as the team-level covariance matrix of the latent variable factor scores and $\mathbf{R}_{\boldsymbol{\eta}}^{L2}$ as a type of factor-specific reliability-based correction matrix. The diagonal elements of the $\mathbf{R}_{\boldsymbol{\eta}}^{L2}$ correction matrix are functions of the respective team-level factor score ($\mathbf{A}^{L2}$), loading ($\boldsymbol{\Lambda}^{L2}$), and indicator reliability ($\boldsymbol{\Omega}_{\boldsymbol{\eta}}$) matrices such that

$$\mathbf{R}_{\boldsymbol{\eta}}^{L2} = \begin{bmatrix} \mathbf{A}_{\eta_1}^{L2}\boldsymbol{\Omega}_{\eta_1}\boldsymbol{\Lambda}_{\eta_1}^{L2} & 0 & \dots & 0 \\ 0 & \mathbf{A}_{\eta_2}^{L2}\boldsymbol{\Omega}_{\eta_2}\boldsymbol{\Lambda}_{\eta_2}^{L2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{A}_{\eta_K}^{L2}\boldsymbol{\Omega}_{\eta_K}\boldsymbol{\Lambda}_{\eta_K}^{L2} \end{bmatrix} \tag{5}$$

The factor score ($\mathbf{A}^{L2}$) and loading ($\boldsymbol{\Lambda}^{L2}$) matrices result directly from the estimation of the factor models. More practically, the product of the factor score and loading matrices for a particular factor (e.g., $\boldsymbol{A}_{\eta_1}^{L2}\boldsymbol{\Lambda}_{\eta_1}^{L2}$ for the first factor) produces a type of measurement model-based estimate of the reliability of the factor scores. In turn, multiplying the covariance between factor scores by the inverse of the reliabilities of the respective factor scores produces a type of disattenuated estimate of the true covariance.

The final term in the correction matrix represents the multivariate indicator reliability matrices ($\boldsymbol{\Omega}_{\eta}$) for the team-level random intercepts of the indicator means. A multivariate indicator reliability matrix for a specific factor (e.g., η_1) can be obtained as

$$\boldsymbol{\Omega}_{\eta_1} = \mathbf{T}_{\eta_1}(\mathbf{T}_{\eta_1} + \mathbf{V}_{\eta_1} / n_1)^{-1} \tag{6}$$

where $\mathbf{T}$ is the team- and $\mathbf{V}$ is the individual-level covariance matrix of the indicators and n_1 is the number of individuals per team. Like the previous correction terms (i.e., $\mathbf{A}^{L2}\boldsymbol{\Lambda}^{L2}$), the multivariate indicator reliability matrices (e.g., $\boldsymbol{\Omega}_{\eta1}$, $\boldsymbol{\Omega}_{\eta2}$, …) introduce further adjustments for the unreliabilities of the predicted indicator means. When indicators are assessed using team-level indicators for a specific construct (e.g., *CPB* in Figure 1), the multivariate reliability matrix corresponding to that factor (e.g., $\boldsymbol{\Omega}_{\eta CPB}$) reduces to an identity matrix.

In unbalanced samples where the number of individuals per team is not constant, we can replace the n_1 term in Equation 6 with the harmonic mean ($\tilde{n}_1$) such that the multivariate indicator reliability matrix for the first factor, for example, becomes

$$\boldsymbol{\Omega}_{\eta_1} = \mathbf{T}_{\eta_1}(\mathbf{T}_{\eta_1} + \mathbf{V}_{\eta_1} / \tilde{n}_1)^{-1} \tag{7}$$

Having corrected the covariance terms the team-level covariance matrix, we can next correct the variance terms of the team-level covariance matrix (diagonal components of ($\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{L2}$)). These terms can be estimated in a similar manner using

$$diag(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{L2}) = (\mathbf{R}_{\boldsymbol{\eta}}^{L2})^{-1} diag(\boldsymbol{\Sigma}_{f_{\boldsymbol{\eta}}}^{L2}) \tag{8}$$

Corrections to the variance terms adjust the variance of the observed factor scores to the model-based estimate of the factor variance or the variance assigned if identifying the scale of the latent variable by fixing the variance (e.g., to unit variance) and freeing all loadings.

In a similar manner, we can also apply Croon-based corrections for the individual-level covariance matrix ($\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{L1}$). A bias-corrected individual-level covariance matrix can be estimated as

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{L1} = (\mathbf{R}_{\boldsymbol{\eta}}^{L1})^{-1} \boldsymbol{\Sigma}_{f_{\boldsymbol{\eta}}}^{L1} (\mathbf{R}_{\boldsymbol{\eta}}^{L1})^{-T} \tag{9}$$

Here, we use $\boldsymbol{\Sigma}_{f\boldsymbol{\eta}}^{L1}$ as the individual-level covariance matrix of the latent variable factor scores and $\mathbf{R}_{\eta}^{L1}$ as a correction matrix. Similar to the corrections for the team-level, the correction matrix draws on the respective individual-level factor score ($\mathbf{A}_{\eta_1}^{L1}$) and loading ($\boldsymbol{\Lambda}_{\eta_1}^{L1}$) matrices such that

$$\mathbf{R}_{\boldsymbol{\eta}}^{L1} = \begin{bmatrix} \mathbf{A}_{\eta_1}^{L1}\boldsymbol{\Lambda}_{\eta_1}^{L1} & 0 & \dots & 0 \\ 0 & \mathbf{A}_{\eta_2}^{L1}\boldsymbol{\Lambda}_{\eta_2}^{L1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{A}_{\eta_K}^{L1}\boldsymbol{\Lambda}_{\eta_K}^{L1} \end{bmatrix} \tag{10}$$

In contrast to the team-level corrections, however, the individual-level corrections do not need to additionally incorporate the reliabilities of the indicator means. The variance terms of the individual-level covariance matrix (diagonal terms of $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{L1}$) can be also estimated using

$$diag(\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{L1}) = (\mathbf{R}_{\boldsymbol{\eta}}^{L1})^{-1} diag(\boldsymbol{\Sigma}_{f_{\boldsymbol{\eta}}}^{L1}) \tag{11}$$

To conceptually summarize, the third step of the Croon method is the crucial correction step of the approach. The initial estimates of the relationships among the latent variables gleaned from the predicted factor-score variance-covariance matrices are biased due to the score uncertainty. For instance, in our working example, the factor score predictions for the leader and team member latent traits under consideration (*CPB*, *PLC*, *CTC*, and *TP*) contain uncertainty or measurement error. Ignoring this error leads to biased estimates of the relationships among the latent variables. However, by using the results of the measurement models (e.g., the factor score matrices, the loading matrices, and the estimated unreliabilities of the factors) as described above, it is possible to account for this uncertainty and reduce bias in the estimated variance-covariance matrix.

In the final step, we use the bias-corrected covariance matrices for each level as sample covariance matrices and estimate the structural path coefficients using the typical sample-based path analysis. For instance, using the corrected team-level covariance matrix ($\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{L2}$), we can estimate the team-level structural coefficient parameters using the path model depicted in Figure (1).

$$\begin{aligned}
\eta_{TP} &= \beta_0^{TP} + b_2\eta_{CTC}^{L2} + b_1\eta_{PLC}^{L2} + c'\eta_{CPB} + \varepsilon_{TP}^{L2} \\
\eta_{CTC}^{L2} &= \beta_0^{CTC} + d_2\eta_{PLC}^{L2} + a_2\eta_{CPB} + \varepsilon_{CTC}^{L2} \\
\eta_{PLC}^{L2} &= \beta_0^{PLC} + a_1\eta_{CPB} + \varepsilon_{PLC}^{L2}
\end{aligned} \tag{12}$$

In this set of equations, we include each of the structural coefficients represented in Figure 1 at the second level (i.e., a_1, a_2, b_1, b_2, c', and d_2). Likewise, using the corrected individual-level covariance matrix ($\boldsymbol{\Sigma}_{\boldsymbol{\eta}}^{L1}$), we can estimate the individual-level structural coefficient parameter (d_1) using the path model

$$\eta_{CTC}^{L1} = \gamma_0^{CTC} + d_1\eta_{PLC}^{L1} + \varepsilon_{CTC}^{L1} \tag{13}$$

In terms of our working example, we estimated the relationships between leader's change-promoting behaviors (*CPB*), individual team member's perceptions of leader charisma (*PLC*), individual team member's commitment to change (*CTC*), and team-level performance from leaders' perspectives (*TP*) using corrected covariance matrices that properly reflected the uncertainty in each latent variable by utilizing estimates from their respective measurement models.

Illustration

We next consider an example analysis that outlines the implementation of the Croon-based estimator for models that integrate top-down and bottom-up effects. Our analyses probe the relationships depicted in Figure 1 and draw on a simulated sample of 33 teams each with 5 team members that were generated on the basis of the correlation matrix provided by Nohe et al. (2013). Code implementing the analyses in *R* using the bcfspa() function is outlined below and the code generating the data and an algorithm implementing Croon's method using the *lavaan* package is available in supplemental materials (Rosseel, 2012). Alternative implementations using the functions directly in *lavaan* package should be available soon.

After downloading the source code for the bcfspa() function, we can first specify the structure of the model using *lavaan* syntax as

```
semodel<- '
   level: 1
   fpcL1 =~ pc1+pc2+pc3
   fctcL1 =~ ctc1+ctc2+ctc3+ctc4
   fctcL1 ~ fpcL1


   level: 2
   fcpbL2 =~ cpb1+cpb2+cpb3+cpb4+cpb5+cpb6
```

```
    fpcL2 =~ pc1+pc2+pc3
    fctcL2 =~ ctc1+ctc2+ctc3+ctc4
    ftpL2 =~ tp1+tp2+tp3+tp4
    ftpL2 ~ fctcL2+fpcL2+fcpbL2
    fctcL2 ~ fpcL2+fcpbL2
    fpcL2 ~ fcpbL2
    '
sem1<-sem(semodel, data=d, cluster="id2")

summary(sem1)
```

The syntax expresses the individual- and team-level reflective common factor models for the latent variables using the '=~' operator. Subsequently, the syntax specifies the regression relationships using the '~' operator.

Once we have specified the form of the measurement and structural models, we can estimate the parameters using the Croon-based method. Using the bcfspa() function, we can estimate the parameters of the model using

```
croon1<-bcfspa(semodel, data=d, cluster="id2",univariate=FALSE)
```

In this syntax, the first argument we specify represents the syntactical representation of the model, the second represents the data (i.e., dataframe), the third represents the cluster or team identifying code, and the fourth represents the method with which we predict indicator means. When the `univariate` argument is true, the function draws on a simpler univariate estimation approach for the indicator reliabilities and when it is false the function draws on multivariate estimates of the indicator reliabilities (see Equation 6 and Kelcey, Cox, & Dong, 2019).

For purposes of comparison, we also estimate the parameters using maximum likelihood and uncorrected factor score path analysis. Maximum likelihood estimates the parameters of each measurement and structural model concurrently. As previously outlined, this approach has shown to yield reduced convergence rates and biased coefficients with small multilevel sample sizes but it represents the predominant approach in practice. It can be implemented using the *lavaan* package as

```
sem1<-sem(semodel, data=d, cluster="id2")
```

As a final comparison, we considered uncorrected factor score path analysis. This approach is a common alternative to maximum likelihood in small sample settings because it tends to be more robust in terms of convergence. The approach is similar to that of Croon's method but draws on the uncorrected covariances among factors to estimate structural relationships. As a result, uncorrected factor score path analysis is known to return biased structural parameter estimates because it neglects the uncertainty of the factor scores. We can implement the approach using the bcfspa() approach as

```
croon_fs<- bcfspa(semodel, data = d, cluster = "id2",
                            univariate=FALSE, uncorrected.cov = TRUE)
```

The syntax specifying the uncorrected factor score path analysis approach follows that of the Croon-based approach. However, the final uncorrected.cov argument allows users to request estimates that avoid the aforementioned Croon-based corrections and so that estimates reduce to the conventional uncorrected factor score path analysis approach.

The comparative results of our illustrative analyses are summarized in Table 1 by path coefficient and estimator. Overall, there is a pattern of nontrivial bias in the parameter estimates for each estimator. Across coefficients, however, the Croon-based estimator returned estimates with the smallest average absolute bias. More specifically, the average absolute bias across coefficients was 0.12 for Croon's method but nearly doubles for the uncorrected factor score path analysis (0.21), and nearly triples for maximum likelihood (0.34). Although these data were purposively chosen to illustrate the potential differences among the estimators, prior research has consistently found that the nature and direction of these discrepancies are indicative of the typical performance of these estimators in small samples (e.g., Kelcey, Cox, & Dong, 2019). Using a simulation, we further probe the findings of past research by examining the extent to which the discrepancies observed in our example analysis among estimators persist across similar and larger samples and a range of conditions for structural equation models that integrate bottom-up and top-down effects.

TABLE 1
Team-level structural model results

Coefficient	True	FS	ML	Croon
a_1	0.24	0.74	–0.02	0.16
a_2	0.18	–0.06	0.37	0.00
c'	0.22	0.18	0.19	0.23
d_2	0.26	–0.02	–0.97	0.00
b_1	0.10	0.00	0.36	0.00
b_2	0.16	0.06	0.10	0.08
Average Abs Bias	--	0.21	0.34	0.12

Note. ML is concurrent estimation of all parameters using maximum likelihood; FS is uncorrected factor score path analysis; Croon is Croon's bias-corrected method

SIMULATION

As previously noted, prior research has suggested that Croon-based estimation in small sample single and multilevel settings often outperforms concurrent maximum likelihood and uncorrected factor score path analysis in terms of convergence, bias, and variance (e.g., Devlieger & Rosseel, 2017; Hayes & Usami, 2019; Kelcey, 2019; Kelcey, Cox, & Dong, 2019). However, these simulations have been largely focused on single-level models and multilevel models that draw solely on lateral and top-down models. Past research has generally excluded the consideration of bottom-up effects and multilevel theories of action that incorporate and integrate bottom-up with lateral and top-down effects. For this reason, we extend prior investigations of the Croon-based estimator by probing its performance in multilevel structural equation models that integrate top-down and bottom-up effects relative to the performances of (uncorrected) factor score path analysis and (concurrent) maximum likelihood.

Facets

Using the multilevel structural equation model depicted in Figure 1 and the correlations reported in Nohe et al. (2013) as the base generating model, we varied six facets in our simulations. The first and second facets were the team-level sample size (n_2) and the individual-level sample size (n_1). Past research has widely found that sample size was a seminal factor in the absolute and relative performance of the estimators (e.g., Croon & van Veldhoven, 2007; Devlieger & Rosseel, 2017). Differences among the estimators have been most pronounced at small to moderate team-level sample sizes but also when the number of individuals per team is small to moderate (e.g., Croon & van Veldhoven, 2007; Kelcey, Cox, & Dong, 2019). Although what constitutes small and moderate sample sizes can be judged relative to a particular field, the complexity of even simple multilevel structural equation models would suggest that samples of less than 100 teams and less than 50 individuals per team falls within that range (e.g., Croon & van Veldhoven, 2007). Prior studies have often considered team-level sample sizes on the order of 50 to 100 and individual-level sample sizes of 10 to 50 per team (e.g., Croon & van Veldhoven, 2007; Li & Beretvas, 2013; Hox et al., 2010). In this study, we extend the range by also considering small team-level sample sizes — such as those found in the literature (e.g., Nohe et al., 2013). Team-level sample size ranged from the original sample size of 33 — used in the Nohe et al. (2013) study and in our example analysis above — to three times that size (99 teams). Likewise, our analyses considered individual-level sample sizes that were particularly small to moderate by the standard of previous simulations but potentially consonant with that of practice (e.g., Hox & Maas, 2001; Nohe et al., 2013). Individual-level sample sizes spanned from the original sample size in Nohe et al. (2013) of about 5 to three times that size (15 individuals per team).

The third facet we considered was the variation in the factor loadings for each factor. Simulations into the performance of Croon-based, uncorrected factor score path analysis, and maximum likelihood estimation have demonstrated that under some conditions the absolute and relative performance of these estimators is influenced by such loadings (e.g., Devlieger & Rosseel, 2017; Kelcey, 2019; Wolf, Harrington, Clark & Miller, 2013). We considered two types of specifications. In the first specification, factors were generated by constraining all indicator loadings to be equal to one (at both the team- and individual-levels). In the second specification, the indicator loadings varied between 0.5 and 1.5.

The fourth facet we probed was cross-level metric (non-)invariance. In multilevel models, multilevel indicators may reflect constructs conceptualized at different levels to varying degrees (e.g., Jak, 2019; Kim, Yoon, Wen, Luo, & Kwok, 2015). When model complexity is high relative to the sample size, a common approach is to reduce the number of parameters in the model by constraining the loadings of an indicator to be equal across levels (e.g., Depaoli & Clifton, 2015; Gonzalez-Roma & Hernandez, 2017; Kelcey, McGinn, & Hill, 2014). Prior research has suggested that under some conditions such cross-level invariance constraints can actually improve convergence without appreciably introducing bias (Kim & Cao, 2015). Such constraints may be particularly relevant for models that include top-down and bottom-up effects because measurement properties, scales, and relationships within teams may be different than those across teams. In our simulation, we examined the performance of the estimators in settings where cross-level metric invariance was and was not maintained. For invariant conditions, the loadings for a single indicator were held equal across levels (but different across indicators) and took values ranging from 0.5 or 1.5. For non-invariant conditions, the loadings for a single indicator opposed each other across levels taking a value of 0.5 at one level and 1.5 at the other level.

Similarly, the fifth facet expanded invariance considerations by examining the impact of different indicator error variance magnitudes across levels. Such modulations help to additionally probe the performance of the estimators relative to the level-specific reliability of the factors. In the invariant case, the indicator error variances at both levels were held to be equal and set to a value of one or two. In our non-invariant case, the indicator error variances opposed each other across levels such that one level was set to one and the other level was set to two (see Table 2 for details).

The final facet we varied was the consistency of the individual-level sample size across clusters. A historical challenge in multilevel settings has been combining varying degrees of information provided by clusters due to unbalanced samples of individuals (e.g., Guittet et al., 2006; Hox & Maas, 2001; Hox et al., 2010; Muthen, 1994; Raudenbush & Bryk, 2002). Prior research has found that unbalanced cluster sizes can have a broad range of detrimental effects on the estimation of parameters in multilevel models and multilevel structural equation models (e.g., Hox et al., 2010; Muthen, 1994; Yuan & Hayash, 2005).

TABLE 2
Simulation results for balanced samples

Sample		Invariance				Convergence			Bias			RMSE		
n_2	n_1	$\Lambda_i = \Lambda_j$	$\Lambda^{L2} = \Lambda^{L1}$	ζ^{L1}	ζ^{L2}	FS	Croon	ML	FS	Croon	ML	FS	Croon	ML
33	5	equal	yes	1	1	.97	.96	.88	0.03	0.02	0.03	0.24	0.31	0.51
45	5	equal	yes	1	1	.99	.99	.95	0.04	0.01	0.01	0.21	0.27	0.37
66	10	equal	yes	1	1	1.00	1.00	1.00	0.04	0.01	0.01	0.15	0.19	0.21
99	15	equal	yes	1	1	1.00	1.00	1.00	0.04	0.00	0.00	0.12	0.14	0.15
33	5	equal	yes	2	2	.71	.70	.55	0.05	0.01	0.03	0.58	0.54	0.84
45	5	equal	yes	2	2	.87	.86	.75	0.05	0.01	0.03	0.29	0.37	0.56
66	10	equal	yes	2	2	.97	.97	.92	0.06	0.00	0.02	0.19	0.27	0.38
99	15	equal	yes	2	2	.99	.99	.98	0.06	0.01	0.01	0.14	0.19	0.23
33	5	mixed	yes	2	2	.63	.62	.46	0.06	0.03	0.06	0.41	0.49	0.80
45	5	mixed	yes	2	2	.73	.72	.63	0.05	0.02	0.04	0.30	0.34	0.66
66	10	mixed	yes	2	2	.89	.89	.86	0.06	0.01	0.04	0.22	0.23	0.60
99	15	mixed	yes	2	2	.95	.95	.95	0.06	0.01	0.01	0.15	0.18	0.21
33	5	mixed	no	1	2	.67	.66	.62	0.05	0.02	0.03	0.55	0.57	0.75
45	5	mixed	no	1	2	.78	.77	.73	0.04	0.02	0.03	0.24	0.27	0.41
66	10	mixed	no	1	2	.89	.89	.89	0.04	0.01	0.01	0.17	0.20	0.29
99	15	mixed	no	1	2	.96	.96	.96	0.04	0.01	0.01	0.13	0.15	0.16
			Average Across Conditions			.87	.87	.82	0.05	0.01	0.02	0.26	0.30	0.46

Note. $\Lambda_i = \Lambda_j$ indicate factor loadings are held equal across indicators whereas $\Lambda^{L2} = \Lambda^{L1}$ indicates that factor loadings are held equal across levels; ζ^{L1} and ζ^{L2} indicate the magnitude of the individual- and team-level error variances for the indicators; ML is concurrent estimation of all parameters using maximum likelihood; FS is uncorrected factor score path analysis; Croon is Croon's bias-corrected method; RMSE indicates the root mean-squared error.

TPM Vol. 27, No. 3, September 2020
477-499 – Special Issue

Despite the documented challenges of unbalanced cluster sizes for a variety of estimators, prior research has not examined the implications of such imbalances on the absolute and relative performance Croon-based estimation. For this reason, we examined the sensitivity of the estimators to different individual sample sizes for each team. We considered two such conditions that we label balanced and unbalanced. In the balanced case, the number of individuals in each team was constant across teams. In the unbalanced case, we randomly omitted 50% of the individual-level sample size for half of the teams while the remaining teams retained the full individual-level sample size specified in the condition (see Table 3).

Results

To evaluate the absolute and relative performance of the estimators, we drew on 1,000 simulated samples in each condition. We summarized the results of the simulation in terms of three criteria: the rate of convergence, bias, and root mean-squared error. Previous investigations have consistently indicated more reliable convergence was a primary advantage of Croon-based estimators when compared to that of maximum likelihood (e.g., Devlieger & Rosseel, 2017). As a result, we evaluated convergence using the proportion of simulated samples for which the estimator converged to a maximum and did not encounter estimation errors (e.g., non-positive definite covariance matrix).

Similarly, prior work has demonstrated that a second core advantage of Croon-based estimation is reduced bias relative to uncorrected factor score path analysis and maximum likelihood in small to moderate samples (e.g., Kelcey, Cox, & Dong, 2019). Bias was evaluated using the average absolute bias of the team-level structural path coefficients

$$\text{Average Absolute Bias} = \frac{\sum_{p=1}^{P} | \bar{B}_p^{L2} - B_p^{L2} |}{P} \tag{14}$$

with $\bar{B}_p^{L2}$ as the average parameter value across all simulated draws for which the estimator converged, B_p^{L2} as true parameter value, and P as the number of team-level path coefficients.

Like the convergence and bias properties, prior research has also suggested that with small to moderate samples, Croon-based estimators are typically less dispersed when compared to their maximum likelihood counterpart but slightly more dispersed than that of uncorrected factor score path analysis (Kelcey, 2019; Kelcey, Cox, & Dong, 2019). That is, the bias reduction associated with the Croon-based corrections is typically bought with an increase in estimator variance relative to the uncorrected factor score path analysis method. To assess this bias-variance tradeoff, we employed the root mean-squared error of the estimators in order to jointly summarize bias-variance tradeoff considerations (e.g., Lüdtke et al., 2011). Root mean-squared error was evaluated using the average root mean-squared error of the team-level structural path coefficients

$$\text{Root mean-squared error} = \sqrt{\frac{\sum_{p=1}^{P} \sum_{i=1}^{I} (\hat{B}_{pi}^{L2} - B_p^{L2})^2}{PI}} \tag{15}$$

with $\hat{B}_{pi}^{L2}$ as the parameter estimate for parameter p in draw i and I as the number of draws.

The results of our simulation are summarized in Tables 2 and 3. Table 2 outlines the performance of the estimators for balanced samples whereas Table 3 outlines the results for unbalanced samples.

TPM Vol. 27, No. 3, September 2020
477-499 – Special Issue

TABLE 3
Simulation results for unbalanced samples

Sample		Invariance				Convergence			Bias			RMSE		
n_2	n_1	$\Lambda_i = \Lambda_j$	$\Lambda^{L2} = \Lambda^{L1}$	ζ^{L1}	ζ^{L2}	FS	Croon	ML	FS	Croon	ML	FS	Croon	ML
33	5	equal	yes	2	2	.06	.01	.01	0.48	0.37	0.56	6.46	1.47	3.30
45	5	equal	yes	2	2	.07	.01	.01	0.28	0.13	0.46	2.90	0.78	4.27
66	10	equal	yes	2	2	.11	.02	.03	0.17	0.15	0.28	3.11	1.12	3.72
99	15	equal	yes	2	2	.19	.04	.07	0.45	0.10	0.25	11.65	0.15	5.20
33	5	mixed	yes	2	2	.06	.01	.01	0.59	0.25	0.26	5.59	1.43	2.83
45	5	mixed	yes	2	2	.07	.01	.01	0.78	0.21	0.36	12.81	1.39	4.65
66	10	mixed	yes	2	2	.11	.02	.03	0.59	0.20	0.41	13.34	0.91	5.27
99	15	mixed	yes	2	2	.19	.04	.08	0.52	0.12	0.25	13.67	0.78	4.96
33	5	equal	yes	2	1	.05	.01	.01	0.29	0.16	0.22	2.36	0.79	2.17
45	5	equal	yes	2	1	.06	.01	.01	0.22	0.13	0.21	22.10	1.28	13.67
66	10	equal	yes	2	1	.12	.04	.04	0.56	0.12	0.37	11.70	0.93	5.02
99	15	equal	yes	2	1	.25	.10	.14	0.15	0.11	0.10	1.17	0.78	3.54
33	5	mixed	no	2	2	.06	.01	.01	3.67	0.32	0.31	42.21	1.75	2.28
45	5	mixed	no	2	2	.06	.01	.02	0.28	0.18	0.36	4.79	0.94	3.00
66	10	mixed	no	2	2	.12	.02	.03	1.59	0.16	0.23	40.39	1.03	5.92
99	15	mixed	no	2	2	.19	.04	.08	0.29	0.12	0.20	72.96	5.02	36.07
			Average Across Conditions			.11	.03	.04	0.68	0.18	0.30	16.70	1.28	6.62

Note: $\Lambda_i = \Lambda_j$ indicate factor loadings are held equal across indicators whereas $\Lambda^{L2} = \Lambda^{L1}$ indicates that factor loadings are held equal across levels; ζ^{L1} and ζ^{L2} indicate the magnitude of the individual- and team-level error variances for the indicators; ML is concurrent estimation of all parameters using maximum likelihood; FS is uncorrected factor score path analysis; Croon is Croon's bias-corrected method; RMSE indicates the root mean-squared error.

Convergence

Overall, uncorrected factor score path analysis and Croon's method returned similar convergence rates and both outperformed the maximum likelihood estimator in balanced samples (Table 2). The most striking result from the simulation was the influence of unbalanced samples on the convergence rates of all the estimators (Table 3). Unbalanced samples had a material impact on both the absolute convergence rates and to a much lesser extent on the relative convergence rates across estimators. For instance, when we held constant the loadings across indicators and levels and set the error variance of the indicators at 2, with 33 teams and 5 individuals per team the convergence rates of the estimators were 0.71 for the uncorrected factor score approach, 0.70 for Croon's method, and 0.55 for maximum likelihood (fifth condition in Table 2). Under those same conditions but with the unbalanced sample, the convergence rates dropped to 0.06 for the uncorrected factor score approach, 0.01 for Croon's method, and 0.01 for maximum likelihood (first condition in Table 3). From an absolute perspective, the average convergence rates across conditions that included balanced samples were 0.88 for Croon's and uncorrected factor score path analysis and 0.83 for maximum likelihood (Table 2). In unbalanced samples, the average convergence rates plummeted to 0.03 for Croon's, 0.11 for uncorrected factor score path analysis, and 0.04 for maximum likelihood (Table 3).

Consistent with prior literature, sample size was also positively associated with convergence. Within the limits of our simulation conditions, increases in either the team- or individual-level sample size tended to improve convergence rates. Modulating the loadings across indicators and levels and error variances across indicators had a comparatively small effect. Across estimators, convergence was minimally impacted when loadings varied across indicators. Similarly, there were slight decreases but little material and consistent movement in terms of convergence rates when loadings varied across levels or when the error variances were increased (Tables 2 and 3).

Bias

Overall, Croon's method returned the smallest level of bias for both balanced and unbalanced samples. In unbalanced samples the disparity among the methods in terms of bias was most pronounced. The average absolute bias across conditions was 0.18 for Croon's, 0.30 for maximum likelihood, and 0.68 for uncorrected factor score path analysis. These differences dissipated with balanced samples — the average absolute bias across balanced conditions was 0.01 for Croon's, 0.02 for maximum likelihood, and 0.05 for uncorrected factor score path analysis.

More generally, the results illustrated a pattern paralleling convergence rates — imbalances in the number of individuals per team across teams materially degraded the accuracy of all the estimators. The rank order of the estimators in terms of bias was preserved but at times unbalanced samples had a differential impact on estimator bias. For example, when we held constant the loadings across indicators and levels and the error variance of the indicators at 2, with 33 teams and 5 individuals per team the biases of the estimators were 0.05 for the uncorrected factor score approach, 0.01 for Croon's method, and 0.03 for maximum likelihood (fifth condition in Table 2). Under those same conditions but with the unbalanced sample, the biases increased to 0.48 for the uncorrected factor score approach, 0.37 for Croon's method, and 0.56 for maximum likelihood (first condition in Table 3).

Sample size entered into a similar role for Croon's method and maximum likelihood but not for the uncorrected method. As sample size increased, the bias of Croon's method and maximum likelihood tended to decrease. The performance of both methods improved with but both were subject to occasional outlying estimates that introduced noise with finite sample simulations. The uncorrected factor score approach, however, did not typically return decreasing bias with larger samples. Both results are consistent with those previously found in the literature in that Croon's and concurrent maximum likelihood tend to improve with size while the bias introduced through uncorrected factor score methods does not disappear with size.

The variability of loadings across indicators had some influence on the bias of the estimators. However, the degree and direction of influence was moderated by estimator and sample balance. For balanced samples, there was little to no difference in bias among estimators when comparing conditions with constant loadings across indicators versus those that varied loadings across indicators. In contrast, with unbalanced samples, mixed loadings were associated with increases and decreases in bias (Table 3). For instance, mixed loadings appeared to consistently amplify bias for uncorrected factor score path analysis while sometimes increasing and sometimes decreasing bias for Croon's method and maximum likelihood.

Differential factor loadings across levels also had inconsistent influence on bias. Under balanced samples, the net impact of cross-level non-invariance was trivial in terms of bias. With unbalanced samples, bias varied as a function of cross-level non-invariance but bias fluctuated in ways that made it difficult to see a clear pattern. In part, these results (and the results more generally), were governed by the poor convergence rates — when a minor proportion of the samples converge, estimates of bias can be heavily influenced by a few extreme parameter estimates.

Root Mean-Squared Error

Overall, uncorrected factor score path analysis returned the lowest root mean-squared error in balanced samples (with Croon's method as a close second) but Croon's method returned the lowest root mean-squared error in unbalanced samples (Tables 2 and 3). To some extent, however, the comparisons of the root mean-squared error (and bias) across estimators in unbalanced samples may be skewed by the non-convergence rates. In many instances, the uncorrected factor score path analysis approach converged but Croon's method and maximum likelihood did not. As a result, many of the estimates under the uncorrected factor score path analysis approach represent challenging samples from which to estimate parameter values whereas the majority of the estimates under the other two approaches are likely to represent less challenging samples.

More generally, the root mean-squared error results followed a pattern similar to that of convergence and bias — it experienced the largest changes when samples were shifted from balanced to unbalanced. Returning to the same example condition as before, when we held constant the loadings across indicators and levels and the error variance of the indicators at 2, with 33 teams and 5 individuals per team the root mean-squared errors were 0.58 for the uncorrected factor score approach, 0.54 for Croon's method, and 0.84 for maximum likelihood (Table 2). Under those same conditions but with the unbalanced sample, the root mean-squared errors increased to 6.46 for the uncorrected factor score approach, 1.47 for Croon's method, and 3.30 for maximum likelihood (Table 3). Likewise, different factor loadings across indicators minimally drove down root mean-squared error whereas non-invariance across levels minimally increased root mean-squared error (Tables 2 and 3).

DISCUSSION

Prior research has acknowledged the foundational and theoretical value of investigating explanatory mechanisms such as macro- and micro-level processes in small to moderate sized studies (e.g., Bodner & Bliese, 2017; Croon & van Veldhoven, 2007; Phelps, Kelcey, Liu, & Jones, 2016; Walton, 2014). At the same time, methodological research has outlined the broad range of limitations that are likely to be encountered when estimating the parameters of multilevel models with small to moderate samples (e.g., Croon & van Veldhoven, 2007; Hox et al., 2010). The intersection of these considerations has led to the research and development of a suite of alternative estimators that attempt to leverage different components of information in order to reduce computationally complexity and improve estimation properties.

In this study, we examined the nature and utility of the Croon-based estimator in the context of multilevel structural equation models that integrated top-down and bottom-up processes. The Croon-based corrections for (co)variances between macro- and micro-level variables centered on the reliabilities of the factors and the reliabilities of the indicator means. Specifically, the correction terms leveraged the factor score and factor loading matrices for each factor and the reliabilities of the indicator means to disattenuate the covariance between latent variables.

The simulation results of this study suggested that many of the advantages of Croon-based estimation relative to maximum likelihood and uncorrected factor score path analysis were retained in multilevel models that integrate top-down and bottom-up effects. For instance, the results suggested that relative to maximum likelihood the Croon-based method returned better convergence rates, marginally smaller levels of bias, and smaller root mean-squared error in balanced samples. Moreover, from a relative standpoint, the advantages of Croon's approach appeared largely insensitive to several factors including sample size and size and equality of factor loadings.

The results regarding sample size balance, however, were more qualified. For example, the simulations demonstrated that even though the Croon-based method converges more regularly and with minimal bias in balanced samples, the performance of both maximum likelihood and Croon's method substantially deteriorated with unbalanced samples. That is, none of the methods considered performed well in an absolute sense when the data were small and unbalanced. Still, in unbalanced samples, while the differences for convergence rates disappeared between Croon's method and maximum likelihood, their differences in bias and root mean-squared error became more pronounced.

The summation of the results still suggests that Croon-based estimation can be useful and effective in small sample multilevel settings. However, the results also introduce caution in its use because the method is not immune to many of the known limitations of other estimators. For instance, although the Croon-based estimator often outperforms maximum likelihood, in an absolute sense its performance can be poor under complicated scenarios such as small unbalanced samples. Further, the results also suggested that the choice of estimator(s) may be context dependent. In balanced samples, for example, although the uncorrected factor score analysis approach maintained higher levels of bias than did Croon's method, the uncorrected factor score analysis approach retained the smallest root mean-squared error and routinely converged at a higher rate.

Collectively, the results suggest that analyses should often take on a type of syndicate approach whereby multiple estimators are applied to a dataset for a given model. Such a strategy may have two useful purposes. First, because each estimator demonstrated significant non-convergence rates across a variety of conditions, convergence under at least one estimator is more likely because each excels under different conditions. For instance, maximum likelihood tends to be the most versatile when there are large samples;

Croon's method tends to perform the best with small to moderate samples; and the uncorrected factor score approach tends to be a type of fallback estimator because it retained the best convergence rates. Moreover, prior research has also supported the potential efficacy of this approach — for instance, Kelcey, Cox, and Dong (2019) found that on the subgroup of samples for which Croon's estimator converged but maximum likelihood did not, Croon's estimator still returned parameter estimates with the same minimal bias. That is, there is evidence that non-convergence of the maximum likelihood estimator has little bearing on whether parameter estimates from Croon's method will incur bias.

Second, given the complementary nature of the estimators and their context-specific advantages, a potentially useful strategy to investigate is to compare the parameter estimates returned by each estimator. That is, the potential use of multiple estimators also introduces the question of the extent to which the similarity of resulting parameter estimates from all three estimators increases the dependability of those estimates. Although the similarity of estimates across estimators does not ensure or bolster their accuracy, their similarity might suggest that the estimates are not sensitive to estimation issues. In contrast, finding evidence of material differences among the estimators suggests that one or more of the estimators may have encountered estimation issues. Past research in different contexts has also used different amalgamations and comparisons of multiple estimators in multilevel models to probe the quality of estimates or the underlying assumptions (e.g., Raudenbush & Bryk, 2002). Future research should probe the degree to which correspondences across the estimators might be indicative of proper estimation convergence or minimal bias.

The results of the study also suggest that additional developments and adjustments to the nature and scope of Croon style corrections may be valuable. For instance, our findings suggest that challenging estimation conditions arise when we encounter a small, moderate, or even a relatively large sample that is unbalanced. Although the convergence performance of Croon's estimator was comparable to that of maximum likelihood, alternative approaches may serve to improve such performance. For example, many of the estimation issues faced in the unbalanced conditions involved non-convergence of one or more of the measurement models. However, some of the non-convergent samples involved the corrected covariance matrix exceeding what is mathematically possible. In the former instances, alternative measurement models or estimation techniques may resolve issues. In these latter instances, alternative estimation approaches may be useful or alternative approaches to the correction terms such as smoothing to prevent non-positive definite covariance matrices may considerably improve the convergence, bias and variance of Croon-based estimation. Similarly, our use of the harmonic mean for the individual-level sample sizes ($\tilde{n}_1$) in unbalanced samples may have contributed to estimation issues. Drawing on alternative approaches — such as smoothing or weighted versions of the multivariate reliability matrix with weights proportional to the team-specific individual-level sample size — may prove more effective. Additional research on the relative roles of the team- and individual-level sample sizes or how their interaction might bolster or undermine the performance of Croon-based estimation may prove useful in selecting estimators or in delineating how planned studies might allocate resources or sample size across levels of the hierarchy.

REFERENCES

Aguinis, H., Boyd, B. K., Pierce, C. A., & Short, J. C. (2011). Walking New Avenues in Management Research Methods and Theories: Bridging Micro and Macro Domains. *Journal of Management*, *37*(2), 395-403. doi:10.1177/0149206310382456

Antonakis, J., Fenley, M., & Liechti, S. (2011). Can charisma be taught? Tests of two interventions. *Academy of Management Learning & Education*, *10*(3), 374-396. doi:10.5465/amle.2010.0012

Asparouhov, T. & Muthen, B. (2007). *Computationally efficient estimation of multilevel high-dimensional latent variable models*. Proceedings of the Joint Statistical Meeting, August, Salt Lake City, Utah.

Bodner, T. E., Bliese, P. D. (2017). Detecting and differentiating the direction of change and intervention effects in randomized trials. *Journal of Applied Psychology*, *103*, 37-53. doi:10.1037/apl0000251

Christ, O., Sibley, C., & Wagner, U. (2012). *Multilevel modeling in personality and social psychology*. Oxford University Press. doi:10.1093/oxfordhb/9780195398991.013.0010

Conger, J. A., Kanungo, R. N., & Menon, S. T. (2000). Charismatic leadership and follower effects. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, *21*(7), 747-767. doi:10.1002/1099-1379(200011)21:7<747::AID-JOB46>3.0.CO;2-J

Cronin, M. A., Weingart, L. R., & Todorova, G. (2011). Dynamics in groups: Are we there yet? *The Academy of Management Annals*, *5*(1), 571-612. doi:10.1080/19416520.2011.590297

Croon, M. (2002). Using predicted latent scores in general latent structure models. In G. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure modeling* (pp. 195-223). Mahwah, NJ: Erlbaum.

Croon, M. A., & van Veldhoven, M. J. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, *12*(1), 45-57. doi:10.1037/1082-989X.12.1.45

DeGroot, T., Kiker, D. S., & Cross, T. C. (2000). A meta-analysis to review organizational outcomes related to charismatic leadership. *Canadian Journal of Administrative Sciences/Revue Canadienne Des Sciences De L'Administration*, *17*(4), 356-372. doi:10.1111/j.1936-4490.2000.tb00234.x

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*(3), 327-351. doi:10.1080/10705511.2014.937849

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher*, *38*(3), 181-199. doi:10.3102/0013189X08331140

Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. *Educational and Psychological Measurement*, *76*(5), 741-770. doi:10.1177/0013164415607618

Devlieger, I., & Rosseel, Y. (2017). Factor score path analysis: An alternative for SEM. *Methodology*, *13*, 31-38. doi:10.1027/1614-2241/a000130

Devlieger, I. & Rosseel, Y. (2019). Multilevel factor score regression. *Multivariate Behavioral Research*. Advance online publication. doi:10.1080/00273171.2019.1661817

Dong, N., Kelcey, B., & Spybrook, J. (2018). Power analyses for moderator effects in three-level cluster randomized trials. *Journal of Experimental Education*, *86*(3), 489-514. doi:10.1080/00220973.2017.1315714

Felfe, J. (2006). Transformational and charismatic leadership: State of research and perspectives. *Zeitschrift Fur Personalpsychologie*, *5*(4), 163-176. doi:10.1026/1617-6391.5.4.163

Gagne, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, *41*(1), 65-83. doi:10.1207/s15327906mbr4101_5

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

Guittet, L., Ravaud, P., & Giraudeau, B. (2006). Planning a cluster randomized trial with unequal cluster sizes: Practical issues involving continuous outcomes. *BMC Medical Research Methodology*, *6*(17), 1-15. doi:10.1186/1471-2288-6-17

Hayes, T., & Usami, S. (2019). Factor score regression in the presence of correlated unique factors. *Educational and Psychological Measurement*, *80*(1), 5-40. doi:10.1177/0013164419854492

Herold, D. M., Fedor, D. B., Caldwell, S., & Liu, Y. (2008). The effects of transformational and change leadership on employees' commitment to a change: A multilevel study. *Journal of Applied Psychology*, *93*(2), 346-357. doi:10.1037/0021-9010.93.2.346

Herscovitch, L., & Meyer, J. P. (2002). Commitment to organizational change: Extension of a three-component model. *Journal of Applied Psychology*, *87*(3), 474-487. doi:10.1037/0021-9010.87.3.474

Hill, D. B. (2006). Theory in applied social psychology: Past mistakes and future hopes. *Theory & Psychology*, *16*(5), 613-640.
doi:10.1177/0959354306067440
Hilpert, J. C., & Marchand, G. C. (2018). Complex systems research in educational psychology: Aligning theory and method. *Educational Psychologist*, *53*(3), 185-202.
doi:10.1080/00461520.2018.1469411
Hox, J. J., & Maas, C. J. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling*, *8*(2), 157-174.
doi:10.1207/S15328007SEM0802_1
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*(2), 157-170.
doi:10.1111/j.1467-9574.2009.00445.x
Hox, J. & McNeish, D. (2020). Small samples in multilevel modeling. In R. van de Schoot & M. Miočević (Eds.), *Small Sample Size Solutions* (pp. 215-225). Routledge Publications.
Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications.* Routledge.
Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Jundt, D. (2005). Teams in organizations: From input-process-output models to IMOI models. *Annual Review of Psychology*, *56*, 517-543.
doi:10.1146/annurev.psych.56.091103.070250
Jak, S. (2019). Cross-level invariance in multilevel factor models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 607-622.
doi:10.1080/10705511.2018.1534205
Jak, S., Oort, F. J., & Dolan, C. V. (2014). Using two-level factor analysis to test for cluster bias in ordinal data. *Multivariate behavioral research*, *49*(6), 544-553.
doi:10.1080/00273171.2014.947353
Kelcey, B. (2019). A robust alternative estimator for small to moderate sample SEM: Bias-corrected factor score path analysis. *Addictive Behaviors*, *94*, 83-98.
doi:10.1016/j.addbeh.2018.10.032
Kelcey, B., Cox, K., & Dong, N. (2019). Croon's bias-corrected factor score path analysis for small-to moderate-sample multilevel structural equation models. *Organizational Research Methods.* Advance online publication.
doi:10.1177/1094428119879758
Kelcey, B., Dong, N., Spybrook, J., & Cox, K. (2017). Statistical power for causally defined indirect effects in group-randomized trials with individual-level mediators. *Journal of Educational and Behavioral Statistics*, *42*(5), 499-530.
doi:10.3102/1076998617695506
Kelcey, B., Dong, N., Spybrook, J., & Shen, Z. (2017). Experimental power for indirect effects in group-randomized studies with group-level mediators. *Multivariate Behavioral Research*, *52*(6), 699-719.
Kelcey, B., Hill, H., & Chin, M. (2019). Teacher mathematical knowledge, instructional quality, and student outcomes: A multilevel mediation analysis. *School Effectiveness & School Improvement*, *30*(4), 398-431.
doi:10.1080/09243453.2019.1570944
Kelcey, B., McGinn, D., & Hill, H. (2014). Approximate measurement invariance in cross-classified rater-mediated assessments. *Frontiers in Psychology*, *5*, 1-13.
doi:10.3389/fpsyg.2014.01469
Kelcey, B., & Phelps, G. (2013). Considerations for designing group randomized trials of professional development with teacher knowledge outcomes. *Educational Evaluation and Policy Analysis*, *35*(3), 370-390.
doi:10.3102/0162373713482766
Kim, E. S., & Cao, C. (2015). Testing group mean differences of latent variables in multilevel data using multiple-group multilevel CFA and multilevel MIMIC modeling. *Multivariate Behavioral Research*, *50*(4), 436-456.
doi:10.1080/00273171.2015.1021447
Kim, E. S., Yoon, M., Wen, Y., Luo, W., & Kwok, O. (2015). Within-level group factorial invariance with multilevel data: Multilevel factor mixture and multilevel MIMIC models. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(4), 603-616.
doi:10.1080/10705511.2014.938217
Kozlowski, S. W., Gully, S. M., Nason, E. R., & Smith, E. M. (1999). Developing adaptive teams: A theory of compilation and performance across levels and time. In D. R. Ilgen & E. D. Pulakos (Eds.), *The Changing Nature of Work Performance: Implications for Staffing, Personnel Actions, and Development* (pp. 240-292). Wiley.
Kozlowski, S. W., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S.W. J. Kozlowski (Eds.), Multilevel theory, research and methods in organizations: Foundations, extensions, and new directions (pp. 3-90). San Francisco, CA: Jossey-Bass.

Lachowicz, M. J., Sterba, S. K., & Preacher, K. J. (2015). Investigating multilevel mediation with fully or partially nested data. *Group Processes & Intergroup Relations*, *18*(3), 274-289. doi:10.1177/1368430214550343

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*(2), 241-264. doi:10.1080/10705511.2013.769391

Loncke, J., Eichelsheim, V. I., Branje, S. J., Buysse, A., Meeus, W. H., & Loeys, T. (2018). Factor score regression with social relations model components: A case study exploring antecedents and consequences of perceived support in families. *Frontiers in Psychology*, *9*, 1699. doi:10.3389/fpsyg.2018.01699

Lu, I. R., Kwan, E., Thomas, D. R., & Cedzynski, M. (2011). Two new methods for estimating structural equation models: An illustration and a comparison with two established methods. *International Journal of Research in Marketing*, *28*(3), 258-268. doi:10.1016/j.ijresmar.2011.03.006

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*(4), 444-467. doi:10.1037/a0024376

Marks, M. A., Sabella, M. J., Burke, C. S., & Zaccaro, S. J. (2002). The impact of cross-training on team effectiveness. *Journal of Applied Psychology*, *87*(1), 3-13. doi:10.1037/0021-9010.87.1.3

Mathieu, J. E., & Chen, G. (2011). The etiology of the multilevel paradigm in management research. *Journal of Management*, *37*(2), 610-641. doi:10.1177/0149206310364663

Mohammed, S., & Dumville, B. C. (2001). Team mental models in a team knowledge framework: Expanding theory and measurement across disciplinary boundaries. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, *22*(2), 89-106. doi:10.1002/job.86

Moss, B., Kelcey, B., & Showers, N. (2014). Classrooms as moderators of development education effectiveness and student achievement. *Community College Review*, *42*, 201-220.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557-585. doi:10.1007/BF02296397

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*(3), 376-398. doi:10.1177/0049124194022003006

Nohe, C., Michaelis, B., Menges, J. I., Zhang, Z., & Sonntag, K. (2013). Charisma and organizational change: A multilevel study of perceived charisma, commitment to change, and team performance. *The Leadership Quarterly*, *24*(2), 378-389. doi:10.1016/j.leaqua.2013.02.001

Phelps, G., Kelcey, B., Liu, S., & Jones, N. (2016). Informing estimates of program effects for studies of mathematics professional development using teacher content knowledge outcomes. *Evaluation Review*, *40*, 383-409.

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

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software*, *48*(2), 1-36.

Rosseel, Y. (2020). Small sample solutions for structural equation modeling. In R. van de Schoot & M. Miočević (Eds.), *Small Sample Size Solutions* (pp. 226-238). Routledge Publications.

Schochet, P. Z. (2011). Do typical RCTs of education interventions have sufficient statistical power for linking impacts on teacher practice and student achievement outcomes? *Journal of Educational and Behavioral Statistics*, *36*(4), 441-471. doi:10.3102/1076998610375840

Smid, S., & Rosseel, Y. (2020). SEM with small samples: two-step modeling and factor score regression versus Bayesian estimation with informative priors. In R. van de Schoot & M. Miočević (Eds.), *Small Sample Size Solutions* (pp. 239-254). Routledge Publications.

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: An examination of the precision of cluster randomized trials funded by the US institute of education sciences. *International Journal of Research & Method in Education*, *39*(3), 255-267. doi:10.1080/1743727X.2016.1150454

Takane, Y., & Hwang, H. (2018). Comparisons among several consistent estimators of structural equation models. *Behaviormetrika*, *45*(1), 157-188.

Walter, F., & Bruch, H. (2009). An affective events model of charismatic leadership behavior: A review, theoretical integration, and research agenda. *Journal of Management*, *35*(6), 1428-1452. doi:10.1177/0149206309342468

Walton, G. M. (2014). The new science of wise psychological interventions. *Current Directions in Psychological Science*, *23*(1), 73-82.
doi:10.1177/0963721413512856

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement*, *73*(6), 913-934.
doi:10.1177/0013164413495237

Wu, J. B., Tsui, A. S., & Kinicki, A. J. (2010). Consequences of differentiated leadership in groups. *Academy of Management Journal*, *53*(1), 90-106.
doi:10.5465/AMJ.2010.48037079

Yuan, K.-H., & Hayash, K. (2005). On Muthen's maximum likelihood for two-level covariance structure models. *Psychometrika*, *70*(1), 147-167.
doi:10.1007/s11336-003-1070-8