Detecting Consensus Emergence in Organizational Multilevel Data: Power Simulations

Jonas W. B. Lang
Department of Personnel Management, Work and Organizational Psychology
Ghent University
Henri Dunantlaan 2
9000 Ghent, Belgium
and
Department of Management
University of Exeter Business School
Rennes Dr, Exeter EX4 4PU, UK
Phone: +32-9-264-6456
Email: jonas.lang@ugent.be

Paul D. Bliese
Darla Moore School of Business
University of South Carolina
1014 Greene Street
Columbia, SC 29208
Phone: 803-777-5957
Email: paul.bliese@moore.sc.edu

J. Malte Runge
Department of Personnel Management, Work and Organizational Psychology
Ghent University
Henri Dunantlaan 2
9000 Ghent, Belgium
Email: malte.runge@ugent.be

in press, Organizational Research Methods

in feature topic "Multilevel Methods and Statistics" (Eds. Eckhardt, Spain, Dionne, Moliterno, & Yammarino)

Abstract

Theories suggest that groups within organizations often develop shared values, beliefs, affect, behaviors or agreed-upon routines; however, researchers rarely study predictors of consensus emergence over time. Recently, a multilevel-methods approach for detecting and studying emergence in organizational field data has been described. This approach—the consensus emergence model—builds on an extended three-level multilevel model. Researchers planning future studies based on the consensus emergence model need to consider (a) sample size characteristics required to detect emergence effects with satisfactory statistical power, and (b) how the distribution of the overall sample size across the levels of the multilevel model influences power. We systematically address both issues by conducting a power simulation for detecting main and moderating effects involving consensus emergence under a variety of typical research scenarios, and provide an R-based tool that readers can use to estimate power. Our discussion focuses on the future use and development of multilevel methods for studying emergence in organizational research.

Keywords: consensus emergence, power analysis, multilevel models

Detecting Consensus Emergence in Organizational Multilevel Data: Power Simulations

In their classic work *The Social Psychology of Organizations*, Daniel Katz and Robert Kahn (Katz & Kahn, 1978) suggested that the essence of an organization is “patterned” human behavior. Building on this idea, organizational research frequently describes and defines groups through attributes such as shared values, affect, common behaviors, or procedures on which the organizational or group members have developed. One important question for organizational research is how the psychological essence of organizations and groups—like shared values and common behavior patterns—develop through interactions among unit-members.

Researchers have used different terms to formally describe patterns of change associated with social interactions. One frequently used term is "emergence" (Cronin & Weingart, 2011; Dansereau, Yammarino, & Kohles, 1999; Felin, Foss, & Ployhart, 2015; Humphrey & Aime, 2014; Kozlowski, 2015; Kozlowski, Chao, Grand, Braun, & Kuljanin, 2013; Morgeson & Hofmann, 1999; Ployhart & Moliterno, 2011). Emergence generally indicates an increase in similarity, agreement, or commonality among unit members that leads to the formation of a shared climate (Ashforth, 1985); however, the term “emergence” can be interpreted more broadly to imply the creation of any new property. Within this broad conceptualization of emergence, an extended range of phenomena are possible including the formation of a state of dissensus – an important process less frequently studied in the organizational literature (Harrison & Klein, 2007; Mathieu, Tannenbaum, Donsbach, & Alliger, 2014). Given these potential definitional ambiguities, we use the narrow term "consensus emergence" to describe increases in shared values, opinions or behaviors over time, and the term “dissensus” to refer to a pattern of decreased similarity in outcome variables. We use the term “emergence” more generally to refer to patterns of change associated with either consensus or dissensus emergence.

One challenge for organizational researchers studying emergence is that the process both

unfolds over time and is simultaneously a multilevel group-phenomenon. This inherent complexity requires a methodological approach that accounts for change over time within higher-level entities, and that captures gradual increases (or decreases) in consensus over time. In contrast, organizational multilevel research has generally been confined to assessing the amount of emergence among unit-members at snapshots in time using cross-sectional multilevel statistics like the intra-class correlation, type 1 (Bliese, 2000). For instance, at first glance a sample of groups with an ICC1 of .02 at time 1, an ICC1 of .15 at time 2, and an ICC1 of .20 at time 3 would appear to be showing a pattern of consensus emergence. Later we describe why it is problematic to interpret raw ICC1 values across time as done here.

Recently, researchers have described an extended three-level multilevel modeling approach—the consensus emergence model (CEM)—that allows researchers to systematically model emergence in the multilevel framework and to study organizational and group characteristics that predict emergence (Lang & Bliese, 2018; Lang, Bliese, & Adler, 2019; Lang, Bliese, & de Voogt, 2018). For instance, an initial study applied the CEM approach to archival data from U.S. Army companies undergoing a major change in core technology and showed that a shared climate of job satisfaction emerged among company members over time (Lang et al., 2018). That is, the finding was focused not on how job satisfaction increased or decreased; rather, the focus was on how soldiers within companies become more similar to each other over time.

The CEM approach can potentially be used to investigate a wide variety of organizational research questions. Nonetheless, three open questions remain for organizational researchers who plan future studies on emergence. First, are questions about sample sizes needed to detect emergence effects with satisfactory statistical power. An initial article on consensus emergence briefly explored this issue by running a power simulation under a typical scenario using 10, 20, and 30 groups. These initial findings suggested that 20 groups were needed (Lang et al., 2018);

however, sample size questions are more complex than captured by Lang et al. (2018) because statistical power may be impacted by combinations of different distributional properties – specifically how observations are distributed across different group sizes, the number of groups and by the overall number of observations. A second related question centers on determining what effect sizes can be detected under different distributional properties, and the third question pertains to how predictors of emergence (e.g., moderation effects) respond to different distributional properties. In this paper, we systematically address these questions by conducting a comprehensive power simulation of emergence effects under a variety of common scenarios. We supplement our simulations with the description of a tool written in the R statistical language that readers can use to conduct power simulations for emergence effects.

An Illustrative Example

Table 1 includes a prototypical dataset with 10 units with 5 members across three time points. The measurements were conducted on a Likert-scale ranging from 1 (strongly disagree) to 5 (strongly agree) with multiple items. The 10 units differ on the basis of a group-level predictor.

A hypothetical example where researchers might encounter a dataset of this type would be perceptions of procedural justice in newly formed work groups that work under pay systems which differ in flexibility. For the purposes of illustration, assume we have access to a continuous pay system rating scale where low values represent low flexibility and high values represent high flexibility. Several researchers have argued that justice perceptions in groups may lead to emergence effects because perceptions of organizational injustice may be contagious (Degoey, 2004; Ehrhart, 2004; A. Li & Cropanzano, 2009; Liao & Rupp, 2005). The underlying idea is that people have a tendency to compare and validate their own emotional reactions to stressful events with others (Barsade, 2002). This validation process may lead to consensus about how events (in particular events related to fairness) should be interpreted. Researchers have long been interested

in contagion effects in organizational field data but statistically showing these types of effects in field data has been challenging so existing data often comes from the laboratory (Ambrose, Harland, & Kulik, 1991; A. Li & Cropanzano, 2009). We use the term "contagion" in this context because there is no explicit goal to form consensus; hence, consensus formation is not deliberate. Researchers may also be interested in consensus in groups where the explicit goal is to come to consensus. As examples, juries deliberate to form a joint opinion (Lang et al., 2019) and teams may need to agree about a negotiation strategy.

Figure 1 shows a prototypical emergence pattern with individual measurements increasingly moving closer to the group trend over time. The figure illustrates a form of heteroscedasticity where the variance among group members is dramatically decreasing. In the CEM, this pattern of variance change is treated as a substantive variable that can be formally tested and predicted. The strong effects in Figure 1 (while illustrative) are not realistic in most organizational data. Figure 2 provides a more realistic pattern for 10 teams – several of the groups (units 4, 5, 6, and 9) appear to show evidence of a consensus emergence pattern. Over time, the responses seem to move closer to the average opinion of the group, but it is not clear how large the effect is nor whether the observed pattern would be statistically significant for the sample as a whole. In the next section, we show how the CEM can be used to formally test the hypothesis that a consensus emergence pattern exists in the data in Table 1 and Figure 2.

The Consensus Emergence Model

We previously noted that a frequently used tool for assessing the presence of consensus in cross-sectional data is the ICC1. Unfortunately, the ICC1 has two severe limitations for modeling consensus emergence over time that make it poorly-suited for detecting consensus emergence patterns in data like the illustrative dataset shown in Table 1.

One limitation of the ICC1 is evident in the formula for the ICC1. This formula is based on

a basic intercept-only multilevel model ($Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}$ where Y is the response, i the unit-member, j the unit, γ_{00} the intercept, u_{0j} the latent group mean and e_{ij} the residual). The formula ($ICC1 = \tau_{00} / [\tau_{00} + \sigma^2]$) defines the ICC1 as the variance, τ_{00}, of the latent group means (u_{0j}) divided by itself plus the variance, σ^2, of the residuals (e_{ij}). In other words, the percentage of variance that group membership explains in the overall variance. The problem is that the ICC1 cannot effectively be used to track changes in emergence over time because two different process can lead to increases in ICC1: Either (a) change in the variance of the latent group means (an increase or reduction in τ_{00}) or (b) change in the amount of similarity that group members show with the group mean (an increase or decreases in σ^2).

Empirical ICC1 values tracked over time frequently fail to show emergence patterns (Allen & O'Neill, 2015)—possibly because changes in τ_{00} along with simultaneous changes in σ^2 work against detecting these types of patterns. A pattern of simultaneous change in τ_{00} and σ^2 applies to the example data in Table 1 and Figure 2. The ICC1 values for these data at T1, T2, and T3 were .11, .02, and .02, respectively. These values do not imply a consensus emergence effect even though Figure 2 seems to provide evidence for an effect of this type.

The second limitation of the ICC1 is that it does not provide a comprehensive modeling framework for studying emergence. To conduct effective research on emergence phenomena, researchers would benefit from a formal statistical test for the presence of emergence, effect size information, and the ability to test for moderators of emergence effects.

The limitations of the ICC1 were the main motivation for the development of the CEM (Lang & Bliese, 2018; Lang et al., 2018). The CEM addresses the limitations of the ICC1 and the need for a modeling framework to formally test for consensus emergence using an extended three-level multilevel model specification. The basic CEM can be written as follows.

$$\text{Level-1: } Y_{tij} = \pi_{0ij} + \pi_{1ij}TIME_t + e_{tij} \quad (1)$$

$$\text{Level-2: } \pi_{0ij} = \beta_{00j} + r_{0ij} \quad (2)$$

$$\pi_{1ij} = \beta_{10j}$$

$$\text{Level-3: } \beta_{00j} = \gamma_{000} + u_{00j} \quad (3)$$

$$\beta_{10j} = \gamma_{100} + u_{10j}$$

$$e_{tij} \sim N(0, \sigma^2\exp[2\delta_1 TIME_t]) \quad (4)$$

$$r_{0ij} \sim N(0, \tau_{00}) \quad (5)$$

$$\begin{pmatrix} u_{00j} \\ u_{10j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \upsilon_{00} & \upsilon_{10} \\ \upsilon_{01} & \upsilon_{11} \end{pmatrix}\right) \quad (6)$$

In these equations, t refers to the measurement occasion, i refers to the unit-member (typically individuals), and j refers to the unit. The model combines the basic model structure of the intercept-only model for the ICC1 with a growth model that accounts for changes (u_{10j}) in the latent unit-means (u_{00j}) along with change in the between group variance over time (captured by υ_{01}, and υ_{11}). The model accounts for the fact that each unit-member provides multiple ratings through the unit-member-specific variance (τ_{00}). The resulting three-level model includes measurements at level-1 nested in unit-members at level-2, and unit-members at level-2 nested in units (i.e., groups) at level-3.

The model is extended in the sense that it goes beyond the standard multilevel models and uses a variance function ($\sigma^2\exp[2\delta_1 TIME_t]$) to model change in the residual variance over time (Culpepper, 2010; Harvey, 1976; Pinheiro & Bates, 2000; Rutemiller & Bowers, 1968). In organizational research and other social sciences, variance functions have usually been included in multilevel models and other regression models to account for potential violations of the homogeneity of residual variance assumption (Bliese & Ployhart, 2002; Culpepper, 2010; Harvey, 1976; Rutemiller & Bowers, 1968; Singer & Willett, 2003). However, research methods

experts have long realized that changes in residual variance functions can also have substantive meaning and can thus be used to gain substantive insights (Goldstein, 2011; Kim & Seltzer, 2011; Pinheiro & Bates, 2000; Raudenbush, 1988). Building on this earlier work, the CEM uses an exponential variance function to account for the gradual increases or decreases in residual variance among unit-members. One advantage of using an exponential variance function is that it yields an effect size estimate, δ_1, that corresponds to an approximate linear increase or decrease in the residual standard deviation σ (square root of the residual variance).[2] That is, when the time variable TIME is coded so that it increases by 1 with each measurement occasion t (e.g., 0, 1, 2,…), δ_1 approximately captures the percent change in the residual standard deviation with each measurement occasion up to about +/- .20 (or 20% change) after which the interpretation is not quite as direct.

In research on consensus emergence, δ_1 is expected to be negative implying a reduction in the residual standard deviation. For instance, when σ is 2.6, δ_1 is -0.06, and *TIME* runs from 0 to 2, the formula for the residual variance at the three measurement occasions would be $\sigma^2 = 2.6^2 \times \exp(2 \times -0.06 \times 0) = 6.76$, $\sigma^2 = 2.6^2 \times \exp(2 \times -0.06 \times 1) = 6$, and $\sigma^2 = 2.6^2 \times \exp(2 \times -0.06 \times 2) = 5.32$, respectively. Taking the square root of the variance, the change pattern in the residual standard deviation is $\sigma = 2.60$, $\sigma = 2.45$, and $\sigma = 2.31$ which is approximately equivalent to a six percent decrease with each measurement occasion (2.45 is 94% of 2.60). To test the significance of δ_1, researchers can use a loglikelihood ratio test that compares a model without the exponential variance function (or $\delta_1 = 0$) with the basic CEM specification shown in Equation 1-6.

The CEM model can be fit in several advanced multilevel modeling software packages like the nlme package (Pinheiro & Bates, 2000) in the R environment (R Core Team, 2018) and Mplus (Lang et al., 2018). Typically, restricted maximum likelihood (REML) estimation is preferred because the δ_1 effect is a component of the variance portion of the model, and REML is

considered more accurate for estimating variance components where fixed-effects remain constant as they do in the CEM model (Pinheiro & Bates, 2000; Singer & Willett, 2003).

While the ICC1 has limitations as a measure of consensus emergence, it has the desirable property of providing information that can be readily interpreted by researchers as the percentage of variance that group membership explains in the overall variance at specific points in time. In some cases, researchers may therefore be interested in translating information from a CEM-based analysis into ICC values for particular points in time to evaluate the degree of overall emergence. This goal can be achieved using the ICCEM coefficient (Lang & Bliese, 2018):

$$\text{ICCEM}_t = \frac{\upsilon_{00} + 2\upsilon_{01}TIME_t + \upsilon_{11}TIME_t^2}{\upsilon_{00} + 2\upsilon_{01}TIME_t + \upsilon_{11}TIME_t^2 + \tau_{00} + \sigma^2\exp(2\delta_1 TIME_t)} \quad (7)$$

In interpreting ICCEM values, researchers can follow existing guidance on interpreting ICC1 values (Bliese, 2000; LeBreton & Senter, 2008). A desirable feature of the ICCEM is that it is model-based and thus more robust and stable than ICC1 values at particular points in time.

While δ_1 values provide an approximate relative measure of change and ICCEM values provide a measure of emergence at particular points in time, researchers may also be interested in an overall measure of effect size or explained variance for emergence effects. A challenge for models that include change in the residual variance is that most approaches for estimating R^2 values in mixed-effects models build on the residuals so that these R^2 values either do not change or they decrease when changes in residual variances are included (see overviews in LaHuis, Hartman, Hakoyama, & Clark, 2014; Rights & Sterba, 2019). Thus, typical R^2 approaches are not useful for extended multilevel mixed-effects models. A solution is to use a generalized R^2 statistic such as the R^2_{LR} (Magee, 1990). This statistic is based on the sample size (N), and the likelihood ratio from a null model with only an intercept (L_0) compared to the model of interest (L_M).

$$R^2_{LR} = 1 - \left(\frac{L_0}{L_M}\right)^{\frac{2}{N}} = 1 - \exp\left(-\frac{2}{N}[\log L_M - \log L_0]\right) \quad (8)$$

R^2_{LR} was originally proposed and discussed in the context of logistic regression analyses, but statisticians later realized that R^2_{LR} represents a generalized R^2 coefficient for a large class of models. For OLS regression, R^2_{LR} is identical to the ordinary OLS R^2. Thus, R^2_{LR} allows researchers to directly compare the explained variance of OLS, mixed-effects models, and complex mixed-effects models like the CEM.

The basic CEM shown in Equation 1-6 above can be relatively easily extended to allow researchers to test for potential effects of moderators on consensus emergence. More specifically, substituting Equation 3 and 4 against the following Equations 9 and 10, respectively, yields a model that tests the effect of a unit-level predictor on emergence.

$$\text{Level-3: } \beta_{00j} = \gamma_{000} + \gamma_{010}(PRED_j) + u_{00j} \quad (9)$$
$$\beta_{10j} = \gamma_{100} + \gamma_{110}(PRED_j) + u_{10j}$$
$$e_{tij} \sim N(0, \sigma^2\exp[2\delta_1 TIME_t + 2\delta_2 PRED_j + 2\delta_3 TIME_t PRED_j]) \quad (10)$$

In using this type of model, researchers should be aware that the CEM is a type of growth model and thus the unit-level predictor should be stable (Ployhart & Kim, 2013; Singer & Willett, 2003). The model is flexible in the sense that the predictor in the model can either be dichotomous or continuous. The interpretation of the δ_2 and δ_3 parameters in the model are analogous to the interpretation of interaction effects in normal regression analyses: δ_2 captures the main effect of the predictor at baseline (when TIME is coded 0 at T1), and δ_3 captures differences in the consensus emergence effect δ_1 for different levels of the predictor.

CEM Analysis of the Illustrative Example

To illustrate the use of the CEM, Table 2 provides the results of a CEM analysis of the illustrative data in Table 1. As shown in Table 2, a consensus emergence effect is present in this

dataset, δ_1 = -0.22. The log-likelihood comparison test between a model without consensus emergence and the CEM suggests that this effect is significant, χ^2 (df = 1, N = 150) = 5.10, p = .02. The significant TIME × PRED interaction effect, δ_3 = 0.44, χ^2(df = 1, N = 150) = 11.22, p < .01, in the residual part of Model 4 in Table 2 also indicates that the environment variable ("PRED" – pay system flexibility in our earlier example) moderates the consensus emergence effect so that the effect over time is stronger when pay system flexibility is high. While hypothetical, these findings suggest that justice contagion is a function of the groups' pay system flexibility: contagion occurs more strongly in the groups with a flexible payment system than in groups with an inflexible system.

Researchers interested in further examining the data can estimate the ICCEM for each time point. The ICCEM estimates from the data are .13, .05, and .04 at TIME = 0, TIME = 1, and TIME =2, respectively, and are similar to the ICC1 values of .11, .02, and .02. The CEM analyses in Table 2 illustrate why the ICCEM (and ICC1) values become smaller. Specifically, in this example the negative covariance term (υ_{01}) leads to a decrease in the between-group variance that is more pronounced than the decrease in the within-group variance lowering ICCEM values. We note that in applied examples with many measurement occasions, the ICCEM will often return values that are easier to interpret. Table 2 also includes the R^2_{LR} estimates and shows that adding the emergence effects increases the amount of explained variance. The final model (model 4) explains 26 percent of the overall variance. Readers interested in running the analyses with the data can consult the dataset and R code provided in Appendix A.

The Power Simulations

Statistical power refers to the long-term probability of detecting a significant effect when the effect is present (Cohen, 1992). A basic convention for statistical power is that it should at least be .80 so a researcher has an 80% chance of detecting the effect. A power analysis

commonly includes several steps. In the first step, a researcher chooses an alpha level (e.g., <.05) and a reasonable *a-priori* expected effect size that seems plausible for the research question on the basis of earlier research and practical considerations (what effect size would be of practical interest). Theoretical considerations may also be of interest but actual information on the magnitude of effect sizes from theory is often rare. After choosing the effect size, the next step is to estimate power for the given effect size. For relatively simple statistical models, it is possible to directly estimate power using formulas (Cohen, Cohen, West, & Aiken, 2003). However, for more complex types of models like multilevel models, power depends on a variety of parameters in the model and their combination so power may more efficiently be estimated using simulations (Bliese & Hanges, 2004; Bolker et al., 2013; Mathieu, Aguinis, Culpepper, & Chen, 2012; Pinheiro & Bates, 2000). In power simulations, the researcher specifies the parameters of the data that he/she expects and then generates a series of datasets from the resulting model using a pseudo random number generator. Because the resulting datasets have been generated from a model for which the underlying parameters are known, power represents the percentage of datasets that return a significant effect when the effect exists.

We conducted two different power simulations. In the first power simulation, we focused on detecting consensus emergence effects. For the CEM, the focus was on the log-likelihood-ratio test (χ^2 -test) which compared a model with a consensus emergence effect to a model without this effect. The second power simulation focused on detecting moderators of consensus emergence. The focus thus was on the log-likelihood-ratio test contrasting a model with a moderator of the consensus emergence effect to a model without this effect. Multilevel literature commonly states that studies require at least 30 to 50 groups (Hox, 2002; Maas & Hox, 2005; Mathieu et al., 2012; Snijders & Bosker, 1993, 1999); however, these recommendations are based on cross-sectional studies and focus on top-down effects and thus apply to situations that

fundamentally differ from the CEM. To gain insights into the requirements for the CEM, we manipulated the total sample size, the sample size at the unit level, and how the observations were distributed among units, unit members, and measurement occasions. Simulation conditions were selected to provide insights into designing field studies – for instance whether including additional measurements might compensate for a smaller number of units.

Method

Table 3 shows the data generating values used for the two simulations. The values in both simulations were based on experience with existing data, effect sizes from initial consensus emergence analyses, and theoretical assumptions. Experience suggests that meaningful consensus emergence effects are typically around -.15. For instance, a reanalysis of a study of group cohesion ratings by psychology students working in teams over six weeks (32 groups / 243 persons / 705 observations) yielded an effect size of δ_1= -0.15 over three time points (Lang & Bliese, 2018). A reanalysis of a dataset on job satisfaction in 34 Army companies measured three times (471 soldiers and a total of 1,351 observations) originally reported by Bliese and Ployhart (2002) revealed an effect size of δ_1= -0.10 (Lang et al., 2018). A more extreme effect size estimate (δ_1= -1.02) was obtained in a reanalysis of Sherif's (1935) classic laboratory study on group norms that included four measurements (see Lang & Bliese, 2018). This group norm study isolated the phenomenon of group norm formation by not providing much additional information to participants: such extreme effect sizes are unlikely in field data.[1]

From a theoretical perspective, it seems reasonable to expect that a typical organizational study could yield a reduction of 15 percent of the variance with each measurement occasion across three measurements (45 percent reduction in the residual variance overall). We therefore used δ_1 = -0.15 as the moderate effect size, δ_1 = -0.08 as a small effect size, and δ_1 = -0.25 as a large effect size across three time points for the simulation. Because we were interested in

comparing the effects of the number of time points on power, we rescaled the time variable when more time points were included in the study by dividing the time variable for the larger number of measurement occasions by 3 so that the effect sizes were equivalent. For instance, with 10 measurement time points 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9 were rescaled to 0, 1/3, 2/3, 1, 4/3, 5/3, 2, 7/3, 8/3, and 3, respectively so that 10 measurement occasions would also have a 45% overall reduction in the moderate effect size condition. All power simulations were conducted in the R (R Core Team, 2018) environment using the nlme package (Pinheiro & Bates, 2000).

Results

The results of the power simulations are provided in Table 4, and Figures 3 and 4. Notice in both figures that power was mostly dependent on the overall number of data points and the effect size. The large effect size of $\delta_1/\delta_3 = -0.25$ yielded sufficient power across all conditions in both sets of simulations. In the first set of simulations focused on detecting consensus emergence, the moderate effect size of $\delta_1 = -0.15$ had acceptable power (> .80) with at least 900 observations no matter how these observations were distributed across units, unit sizes, and time points (see Figure 3). While a value of 900 seems large, consider that 30 groups with ten group members over three measurement occasions produces 900 observations. The small effect size, in contrast, required 3,000 observations to consistently yield sufficient power.

In the second set of simulations focused on the power to detect moderators of consensus emergence, the moderate effect size also yielded sufficient power with 900 observations in three of the four conditions (see Figure 4). However, in the fourth condition with a large number of unit-members, and a large number of time points, power was insufficient. The reason is that just 6 groups are not adequate to effectively account for sample size variation of the moderator. In contrast, 10 units were sufficient to generate acceptable power with a large number of measurement occasions (15) and a large number of unit members (10). Again, the small effect

size required 3000 observations to consistently yield sufficient power.

Power Simulation Tool

The power simulations cover a set of typical scenarios in organizational research. However, researchers may want to do their own power simulation on the basis of more specific expectations. Appendix B and C provide R code allowing researchers to conduct simulations of this type. The R functions include model components that a researcher can specify *a-priori* and additionally the setting "tscale" allows researchers to more easily study the impact of varying numbers of time points without changing the metric of the values entered into the model. tscale simply rescales the time variable to make scenarios comparable. For instance, if a researcher wants to compare a scenario with 9 or 3 time points, he/she enters tscale=3 in the model with 9 time points so that all other data generating values for the simulation are equivalent.

Discussion

Emergence represents a multilevel process describing how lower-level units change over time to form characteristics of higher-level units. The concept of emergence plays several roles in theory development and in advancing research. First, emergence often provides the theoretical foundation for aggregating responses to higher-levels and conducting research using higher-level constructs. That is, demonstrating that emergent processes occur for specific constructs represents an important aspect of the multilevel construct validation process that help justifies aggregation.

Second, on a related note, examining emergence via the CEM can provide a deeper understanding of existing unit-level constructs. For instance, unit-level constructs like justice climate or safety climate (e.g., Zohar, 2010) are well-established predictors of organizational outcomes. As a field, however, organizational research does not currently know if justice climates and safety climates develop through emergent processes where unit members become more similar over time with relatively little mean change across units, or whether safety climates

develop because units becoming more extreme in term of mean differences. That is, a significant ICC value at one time point does not provide information about the process that produced it.

Third, the ability to model predictors provides a way to develop and test procedures to predict emergence. Again, using the example of safety climate, researchers and practitioners could determine whether a specific training program causes units to more quickly develop shared safety climates. A study of this nature would be interested in mean change over time, but in addition to mean change, it would be important to examine patterns of emergence among unit members with an eye towards enhancing consensus. As another example, research could also be framed around understanding why some group members become more divergent over time and incorporate predictors of these divergent patterns. Work in this area is important because a lack of consensus among group members is presumably an index of a poorly functioning team.

In the end, studies of emergence potentially have much to offer in terms of theory development. Like other areas of research, studies of emergence need to have acceptable statistical power to help advance knowledge. We used simulation studies to examine how the statistical power of the CEM was related to differences in overall sample sizes, the number of groups, group size, and the number of time points. In addition, we provide R-based tools that researchers can use to conduct power simulations. Our results suggest that the power of the CEM mainly depends on the overall number of data points and the effect size. The distribution of data points across units, time, and unit-members generally appears to have a limited influence on the results. An important exception is that power to detect moderation is substantially lowered by having a small number of units in combination with a high number of measurement occasions and unit members. In this scenario, the large number of measurement occasions and unit members cannot compensate for the lack of information related to the moderator due to the small number of units.

Overall, the results suggest that in many situations a relatively small number of groups may be sufficient when there are either many group members or many measurements for detecting consensus emergence. These results are encouraging in suggesting that datasets that include a limited number of groups can potentially be used to generate novel and interesting insights when frequent measurements are possible. The results may also open the door for studies that track a small number of groups for an extended period of time to see whether group members come together to gain insights on group functioning. An initial example for a study of this type is a diary study that tracked a small number of groups of archeologists on a field mission over several weeks (Lang et al., 2018). The results also suggest that the sample size requirements for detecting emergence are lower than for other types of multilevel effects like, for instance, cross-level moderation effects (Maas & Hox, 2005; Mathieu et al., 2012; Snijders & Bosker, 1993, 1999).

Limitations

We note several limitations of this work. One limitation relates to the nature of the model on which we focused in this article—the consensus emergence model (CEM). The CEM fundamentally assumes that changes in the residual variance convey important and relevant information about emergence, and that patterns of variance change are manifest reflections of changes in group climate. We see the CEM as complementing other approaches and note that the literature on emergence has described several other types of emergence phenomena and alternative methods to study these complex phenomena. Other models to study emergence include qualitative research methods (Gehman, Trevino, & Garud, 2013), computational models that simulate complex emergence processes to gain insights into plausible explanations for empirical patterns (Kozlowski et al., 2013), and network models (Fowler & Christakis, 2008). A review of these methods is beyond the scope of this article but interested readers may examine reviews and overviews of these methods (Kozlowski et al., 2013; Lang et al., 2018). We believe

that these alternative methods will complement the multilevel approach for studying emergence.

We also note that a limitation of the CEM is that it—like all linear mixed-effect multilevel models—assumes normally distributed residuals. This assumption does not mean that the dependent variable itself needs to be normally distributed; rather, the assumption is that the residuals of the model are approximately normally distributed after accounting for all model components. In practice, the assumption of normally distributed residuals implies that users of the CEM should be cautious because heavily skewed data with strong floor and ceiling effects can violate the assumption of normally distributed residuals. A recommended strategy is to examine the residuals using graphical model checking procedures (Pinheiro & Bates, 2000).

Finally, a specific limitation of our study is the fact that we only examined a limited set of conditions so our power simulations will likely not cover some situations that researchers will face in their research. Nonetheless, we attempted to test scenarios that reflected common data characteristics with respect to factors such as group size, the number of measurement occasions, and the number of groups. In addition, the R code in the Appendices B and C can be used to conduct power simulation studies tailored to the specific attributes of the research setting.

Future Directions

One area for future research could be to extend the CEM to study more complex phenomena. One possible extension is to include more complex types of change. The analyses we considered in this study were limited to linear change in both latent group means and consensus. Change, however, may show forms such as a quadratic emergence trajectory where a group first becomes more homogenous and then more heterogeneous (e.g., Tuckman & Jensen, 1977). Change in consensus could also be discontinuous (Bliese & Lang, 2016; Singer & Willett, 2003) because a catalyst event could occur in a group and lead to a pattern of either consensus or dissensus. These more complex consensus change models can be specified relatively easily using

the procedures described in this paper by adding the respective change terms both in the fixed effects part (e.g., $Y_{tij} = \pi_{0ij} + \pi_{1ij}TIME_t + \pi_{2ij}TIME_t^2 + e_{tij}$) and the variance change part (e.g., $\sigma^2\exp[2\delta_1 TIME_t]\exp[2\delta_1 TIME_t^2]$) of the model. However, the interpretation of these more complex models can be challenging and future research needs to study the statistical power for detecting more complex consensus change effects.

Another potential extension of the CEM is to include additional complexity in the consensus change part of the model. For instance, it is possible to add an additional level of analysis to the CEM to account for measurement error (Lang et al., 2018). This approach requires one to use a somewhat different parametrization of variance function models (Goldstein, 2005, 2011) that is not as easy to interpret as the parametrization with the exponential variance functions we generally recommend, but can be fit in almost all multilevel software packages. To implement this approach, one adds time as a predictor centered at the end of the observation period [$TIME_t - \max(TIME_t)$] at the person level (now Level 2 because of the additional level of nesting) and specifies that the time variable is uncorrelated with the intercept. The advantage of this alternative model is that it allows both the intercept and the slope to vary across individuals with the assumption that both are from a normal distribution (random effects). A somewhat similar approach pursued by researchers is to directly add a random intercept effect to the exponential variance specification for the residual variance (Culpepper, 2010; X. Li & Hedeker, 2012). These types of models typically do not include random slopes for time like the CEM but have the advantage that they can even be fit with a joint random effects distribution so that the random error variability can be correlated with the other random effects in the model. Although interesting, these more complex models are frequently not easy to fit for three-level structures and may run into convergence problems unless the sample size is very large (X. Li & Hedeker, 2012; Nestler, Geukes, & Back, 2018). Ultimately, it is important for researchers to carefully

balance model complexity and model parsimony (Bates, Kliegl, Vasishth, & Baayen, 2015; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017).

From a theoretical perspective, it may be interesting to use the CEM to capture more complex forms of emergence. One way this can already be done within the framework of the CEM is to add dichotomous predictors (e.g., leaders vs. non-leaders or ethnic minority vs. non-ethnic minority, or male vs. female) that separate subgroups within groups from each other. For instance, a recent article provides an illustration of how to test consensus emergence among minority and majority group members in mock juries (Lang et al., 2019). We anticipate opportunities to extend the model by looking at other predictors associated with group members.

A set of final questions for future research center on how the CEM can be combined with other established techniques in the multilevel literature like multi-membership models (Cafri, Hedeker, & Aarons, 2015) or mediation models (MacKinnon, Fairchild, & Fritz, 2007). Multi-membership models allow persons to be members of multiple groups. We are not aware of work examining how multi-membership affect emergence. Mediation models are frequently discussed in the literature but it is not clear how the approach could be extended to emergence models. Even with this limitation, however, we note that the ability to include group-level predictors provides a potentially powerful way to examine mechanisms that lead to emergence. Indeed, researchers have argued that one possible approach for testing mediation theories in practice is to manipulate both the predictor and the mediator (Spencer, Zanna, & Fong, 2005). One way to study mediation using the CEM is therefore to experimentally manipulate both the predictor and the mediator and to then test both using a CEM model with a dichotomous predictor for the experimental condition.

References

Allen, N. J., & O'Neill, T. A. (2015). The trajectory of emergence of shared group-level constructs. *Small Group Research*, *46*, 352–390. https://doi.org/10.1177/1046496415584973

Ambrose, M. L., Harland, L. K., & Kulik, C. T. (1991). Influence of social comparisons on perceptions of organizational fairness. *Journal of Applied Psychology*, *76*, 239–246. https://doi.org/10.1037/0021-9010.90.2.242

Ashforth, B. E. (1985). Climate formation: Issues and extensions. *Academy of Management Review*, *10*, 837–847. https://doi.org/10.2307/258051

Barsade, S. G. (2002). The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly*, *47*, 644–675. https://doi.org/10.2307/3094912

Bates, D. M., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious Mixed Models, 1–27. Retrieved from http://arxiv.org/abs/1506.04967

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

Bliese, P. D., & Hanges, P. J. (2004). Being both too liberal and too conservative: The perils of treating grouped data as though they were independent. *Organizational Research Methods*, *7*, 400–417. https://doi.org/10.1177/1094428104268542

Bliese, P. D., & Lang, J. W. B. (2016). Understanding relative and absolute change in discontinuous growth models. *Organizational Research Methods*, *19*(4), 562–592. https://doi.org/10.1177/1094428116633502

Bliese, P. D., & Ployhart, R. E. (2002). Growth modeling using random coefficient models: Model building, testing, and illustrations. *Organizational Research Methods*, *5*, 362–387.

https://doi.org/10.1177/109442802237116

Bolker, B. M., Gardner, B., Maunder, M., Berg, C. W., Brooks, M., Comita, L., … Zipkin, E. (2013). Strategies for fitting nonlinear ecological models in R, AD Model Builder, and BUGS. *Methods in Ecology and Evolution*, *4*(6), 501–512. https://doi.org/10.1111/2041-210X.12044

Cafri, G., Hedeker, D., & Aarons, G. A. (2015). An introduction and integration of cross-classified, multiple membership, and dynamic group random-effects models. *Psychological Methods*, *20*(4), 407–421. https://doi.org/10.1037/met0000043

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Erlbaum.

Cronin, M. A., & Weingart, L. R. (2011). Dynamics in groups: Are we there yet ? *Academy of Management Annals*, *5*, 37–41. https://doi.org/10.1080/19416520.2011.590297

Culpepper, S. A. (2010). Studying individual differences in predictability With gamma regression and nonlinear multilevel models. *Multivariate Behavioral Research*, *45*(1), 153–185. https://doi.org/10.1080/00273170903504885

Dansereau, F., Yammarino, F. J., & Kohles, J. C. (1999). Multiple levels of analysis from a longitudinal perspective: Some implications for theory building. *Academy of Management Review*, *24*, 346–357. https://doi.org/10.2307/259086

Degoey, P. (2004). Contagious justice: Exploring the social construction of justice in organizations. *Research in Organizational Behavior*, *22*, 51–102. https://doi.org/10.1016/S0191-3085(00)22003-0

Ehrhart, M. G. (2004). Leadership and procedural justice climate as antecedents of unit-level

organizational citizenship behavior. *Personnel Psychology*, *57*, 61–94. https://doi.org/10.1111/j.1744-6570.2004.tb02484.x

Felin, T., Foss, N. J., & Ployhart, R. E. (2015). The microfoundations movement in strategy and organization theory. *Academy of Management Annals*, *9*, 575–632. https://doi.org/10.1080/19416520.2015.1007651

Fowler, J. H., & Christakis, N. A. (2008). Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *BMJ (Clinical Research Ed.)*, *337*, a2338. https://doi.org/10.1136/bmj.a2338

Gehman, J., Trevino, L., & Garud, R. (2013). Values work: A process study of the emergence and performance of organizational values practices. *Academy of Management Journal*, *56*, 84–112. https://doi.org/10.5465/amj.2010.0628

Goldstein, H. (2005). Heteroscedasticity and complex variation. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia for statistics in behavioral science (Volume 2)* (pp. 790–795). Chichester, UK: John Wiley & Sons.

Goldstein, H. (2011). *Multilevel statistical models (4th ed.)*. *Book*. Chichester, UK: Wiley. https://doi.org/10.2307/1534624

Harrison, D. A., & Klein, K. J. (2007). What's the difference? Diversity constructs as separation, variety, or disparity in organizations. *Academy of Management Review*, *32*, 1199–1228. https://doi.org/10.5465/AMR.2007.26586096

Harvey, A. C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica*, *44*(3), 461. https://doi.org/10.2307/1913974

Hox, J. (2002). *Multilevel analysis: Thechniques and applications*. Mahwah, NJ: Erlbaum.

Humphrey, S. E., & Aime, F. (2014). Team microdynamics: Toward an organizing approach to teamwork. *The Academy of Management Annals*, *8*, 443–503.

https://doi.org/10.1080/19416520.2014.904140

Katz, D., & Kahn, R. L. (1978). *The social psychology of organizations*. New York, NY, US: Wiley.

Kim, J., & Seltzer, M. (2011). Examining heterogeneity in residual variance to detect differential response to treatments. *Psychological Methods*, *16*, 192–208. https://doi.org/10.1037/a0022656

Kozlowski, S. W. J. (2015). Advancing research on team process dynamics: Theoretical, methodological, and measurement considerations. *Organizational Psychology Review*, *5*, 270–299. https://doi.org/10.1177/2041386610376255

Kozlowski, S. W. J., Chao, G. T., Grand, J. A., Braun, M. T., & Kuljanin, G. (2013). Advancing multilevel research design: Capturing the dynamics of emergence. *Organizational Research Methods*, *16*, 581–615. https://doi.org/10.1177/1094428113493119

LaHuis, D. M., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2014). Explained Variance Measures for Multilevel Models. *Organizational Research Methods*, *17*(4), 433–451. https://doi.org/10.1177/1094428114541701

Lang, J. W. B., & Bliese, P. D. (2018). A temporal perspective on emergence: Using three-level mixed effects models to track consensus emergence in groups. In S. E. Humphrey & J. M. LeBreton (Eds.), *The Handbook for multilevel theory, measurement, and analysis* (pp. 519–540). Washington, DC: APA.

Lang, J. W. B., Bliese, P. D., & Adler, A. B. (2019). Opening the black box: A multilevel framework for studying group processes. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.1177/2515245918823722

Lang, J. W. B., Bliese, P. D., & de Voogt, A. (2018). Modeling consensus emergence in groups using longitudinal multilevel models. *Personnel Psychology*, *71*, 255–281.

https://doi.org/10.1111/peps.12260

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 Questions and Interrater Agreement. *Organizational Research Methods*, *11*, 815–852. https://doi.org/10.1177/1094428106296642

Li, A., & Cropanzano, R. (2009). Fairness at the group level: Justice climate and intraunit justice climate. *Journal of Management*. https://doi.org/10.1177/0149206308330557

Li, X., & Hedeker, D. (2012). A three-level mixed-effects location scale model with an application to ecological momentary assessment data. *Statistics in Medicine*, *31*(26), 3192–3210. https://doi.org/10.1002/sim.5393

Liao, H., & Rupp, D. E. (2005). The impact of justice climate and justice orientation on work outcomes: A cross-level multifoci framework. *Journal of Applied Psychology*, *90*, 242-56. https://doi.org/10.1037/0021-9010.90.2.242

Maas, C. J., & Hox, J. (2005). Sufficient sample sizes for multilevel modeling. *Journal of Research Methods for the Behavioral and Social Sciences*, *1*, 86–92. https://doi.org/10.1027/1614-1881.1.3.86

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, *58*(1), 593–614. https://doi.org/10.1146/annurev.psych.58.110405.085542

Magee, L. (1990). R2 measures based on wald and likelihood ratio joint significance tests. *The American Statistician*, *44*, 250–253. https://doi.org/10.1080/00031305.1990.10475731

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. https://doi.org/10.1037/a0028380

Mathieu, J. E., Tannenbaum, S. I., Donsbach, J. S., & Alliger, G. M. (2014). A review and integration of team composition models: Moving toward a dynamic and temporal

framework. *Journal of Management*, *40*, 130–160. https://doi.org/10.1177/0149206313503014

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Morgeson, F., & Hofmann, D. (1999). The structure and function of collective constructs: Implications for multilevel research and theory development. *Academy of Management Review*, *24*, 249–265. https://doi.org/10.5465/AMR.1999.1893935.

Nestler, S., Geukes, K., & Back, M. D. (2018). Modeling Intraindividual Variability in Three-Level Multilevel Models. *Methodology*, *14*(3), 95–108. https://doi.org/10.1027/1614-2241/a000150

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York, NY, NY: Springer.

Ployhart, R. E., & Kim, Y. (2013). Dynamic longitudinal growth modeling. In J. M. Cortina & R. S. Landis (Eds.), *SIOP organizational frontiers series. Modern research methods for the study of behavior in organizations* (pp. 63–98). New York, NY, US: Routledge/Taylor & Francis Group.

Ployhart, R. E., & Moliterno, T. P. (2011). Emergence of the human capital resource: A multilevel model. *Academy of Management Review*, *36*, 127–150. https://doi.org/10.5465/amr.2011.55662569

R Core Team. (2018). R: A language and environment for statistical computing [Version 3.5.0]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Raudenbush, S. W. (1988). Estimating change in dispersion. *Journal of Educational and*

Behavioral Statistics, *13*, 148–171. https://doi.org/10.3102/10769986013002148

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, *24*(3), 309–338. https://doi.org/10.1037/met0000184

Rutemiller, H. C., & Bowers, D. A. (1968). Estimation in a heteroscedastic regression model. *Journal of the American Statistical Association*, *63*(322), 552–557. https://doi.org/10.1080/01621459.1968.11009274

Sherif, M. (1935). A study of some social factors in perception. *Archives of Psychology*, *27*, 1–60.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurence*. New York, NY: Oxford University Press.

Snijders, T. A. B., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational and Behavioral Statistics*, *18*, 237–259. https://doi.org/10.3102/10769986018003237

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis*. London, UK: Sage.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, *89*(6), 845–851. https://doi.org/10.1037/0022-3514.89.6.845

Tuckman, B. W., & Jensen, M. A. C. (1977). Stages of small-group development revisited. *Group & Organization Management*, *2*(4), 419–427. https://doi.org/10.1177/105960117700200404

Zohar, D. (2010). Thirty years of safety climate research: Reflections and future directions. *Accident Analysis & Prevention*, *42*(5), 1517–1522.

https://doi.org/10.1016/j.aap.2009.12.019

Footnotes

[1]Both the job satisfaction data and the Sherif data are available in the multilevel package for R (Bliese, 2016).

[2]The exponential specification, $\sigma^2 \exp[2\delta_1 TIME_t]$ is a shortened way to write the equivalent form, $(\sigma \exp[\delta_1 TIME_t])^2$, using one less pair of parentheses.

Table 1
Example Dataset

		Unit									
Unit member	TIME	1	2	3	4	5	6	7	8	9	10
1	0	3.5	2.5	3.3	4.0	3.9	3.8	2.9	3.0	1.5	3.3
1	1	3.1	3.1	3.1	3.2	3.0	3.5	2.8	3.3	2.8	2.7
1	2	3.5	3.0	3.2	3.7	3.4	2.9	2.7	3.4	3.1	3.3
2	0	4.5	2.8	3.2	2.0	4.1	3.4	3.2	4.3	2.5	2.3
2	1	4.0	2.8	2.5	2.8	2.8	3.4	3.4	4.1	2.7	2.3
2	2	4.2	4.3	2.5	2.9	2.9	3.0	3.2	4.0	3.4	3.0
3	0	3.1	2.6	3.3	3.4	2.6	3.2	2.9	2.3	3.0	2.3
3	1	3.5	3.6	3.4	2.9	3.4	2.1	3.2	3.0	2.3	3.2
3	2	3.0	3.0	3.6	3.4	3.0	2.9	3.4	2.4	3.0	2.8
4	0	2.6	3.1	3.2	3.3	3.2	2.7	2.7	3.6	3.0	3.3
4	1	3.1	3.6	2.7	3.7	4.0	2.9	2.4	3.4	3.0	3.0
4	2	2.7	3.5	3.0	3.4	3.9	2.9	2.4	3.5	2.8	3.1
5	0	3.6	2.6	3.3	3.4	2.9	3.9	3.0	3.6	3.0	3.0
5	1	3.2	2.3	2.6	2.5	2.7	3.0	3.0	2.9	2.5	3.6
5	2	3.5	3.7	2.8	3.0	2.9	3.5	2.8	3.1	2.9	2.7
Predictor		0.2	2.0	1.0	-0.3	-1.0	-0.3	-0.2	0.1	0.1	0.4

Table 2
Baseline Model (Model 1), Consensus Emergence Model (Model 2), and Test of a Moderator of Consensus Emergence (Model 3 and Model 4) Fitted to the Example Dataset

Parameters	Model 1	Model 2	Model 3	Model 4
Intercept, γ_{000}	3.08	3.07	3.09	3.08
TIME, γ_{100}	0.03	0.04	0.02	0.02
PRED, γ_{010}	—	—	-0.15	-0.14
TIME × PRED, γ_{110}	—	—	0.11	0.08
Unit intercept variance, υ_{00}	0.05	0.04	0.03	0.04
Unit slope variance, υ_{11}	0.01	0.01	0.004	0.01
Unit covariance, υ_{01}	-0.03	-0.02	-0.01	-0.02
Unit-member variance, τ_{00}	0.09	0.09	0.09	0.10
Residual variance, σ^2	0.14	0.20	0.21	0.20
TIME, δ_1	—	-0.22	-0.27	-0.35
PRED, δ_2	—	—	0.09	-0.37
TIME × PRED, δ_3	—	—	—	0.44
logLik	-97.92	-95.37	-96.28	-90.67
df	7	8	11	12
χ^2 vs. previous model		5.10*		11.22*
R^2_{LR}	.15	.18	.21	.26

Note. 10 units with 5 unit-members measured at 3 measurement occasions (150 observations).
* $p < .05$

Table 3
Values Used in the Simulations

Used in the simulation	Simulation 1	Simulation 2
Intercept, γ_{000}	3.00	[a]
TIME, γ_{100} (slope)	0.01	[a]
PRED, γ_{010}		0.01
TIME × PRED, γ_{110}		0.01
Unit intercept variance, υ_{00}	0.05	[a]
Unit slope variance, υ_{11}	0.005	[a]
Unit covariance, υ_{01}	-0.005	[a]
Unit-member variance, τ_{00}	0.10	[a]
Residual variance, σ^2	0.20	[a]
Total observations	450, 900, 1500, 3000	[a]
Distribution of total observations	30/60/100/200 × 5 × 3	[a]
(units * members/unit * timepoints)	15/30/50/100 × 10 × 3	
	15/30/50/100 × 3 × 10	
	3/6/10/20 × 10 × 15	
Effect size δ_1 (TIME)	0, -0.08, -0.15, -0.25	[a]
Effect size δ_2 (PRED)		0, -0.08, -0.15, -0.25
Effect size δ_3 (TIME × PRED)		0, -0.08, -0.15, -0.25

Note. [a]Like in Simulation 1.

Table 4
Results of the Simulations

Units	Members in each unit	Timepoints	Total *N*	Power in simulation 1				Power in simulation 2			
				δ_1 = .00	δ_1 = -.08	δ_1 = -.15	δ_1 = -.25	δ_3 = .00	δ_3 = -.08	δ_3 = -.15	δ_3 = -.25
30	5	3	450	0.051	0.290	0.723	0.986	0.052	0.277	0.687	0.960
60	5	3	900	0.048	0.512	0.949	1.000	0.050	0.493	0.930	0.999
100	5	3	1,500	0.053	0.729	0.996	1.000	0.050	0.715	0.995	1.000
200	5	3	3,000	0.051	0.950	1.000	1.000	0.049	0.947	1.000	1.000
15	10	3	450	0.055	0.299	0.747	0.987	0.052	0.280	0.670	0.941
30	10	3	900	0.052	0.520	0.964	1.000	0.051	0.503	0.928	0.999
50	10	3	1,500	0.054	0.747	0.998	1.000	0.048	0.727	0.994	1.000
100	10	3	3,000	0.052	0.957	1.000	1.000	0.052	0.951	1.000	1.000
15	3	10	450	0.052	0.284	0.717	0.988	0.049	0.254	0.660	0.946
30	3	10	900	0.048	0.491	0.952	1.000	0.050	0.470	0.920	0.999
50	3	10	1,500	0.052	0.703	0.996	1.000	0.049	0.683	0.991	1.000
100	3	10	3,000	0.052	0.942	1.000	1.000	0.047	0.934	1.000	1.000
3	10	15	450	0.051	0.285	0.736	0.990	0.047	0.203	0.457	0.707
6	10	15	900	0.049	0.491	0.957	1.000	0.049	0.402	0.803	0.958
10	10	15	1,500	0.046	0.705	0.998	1.000	0.047	0.628	0.953	0.998
20	10	15	3,000	0.049	0.949	1.000	1.000	0.049	0.900	0.999	1.000

Note. Convergence rate was 99.73% in simulation 1 and 99.66% in simulation 2. 10,000 simulation runs for both simulations.

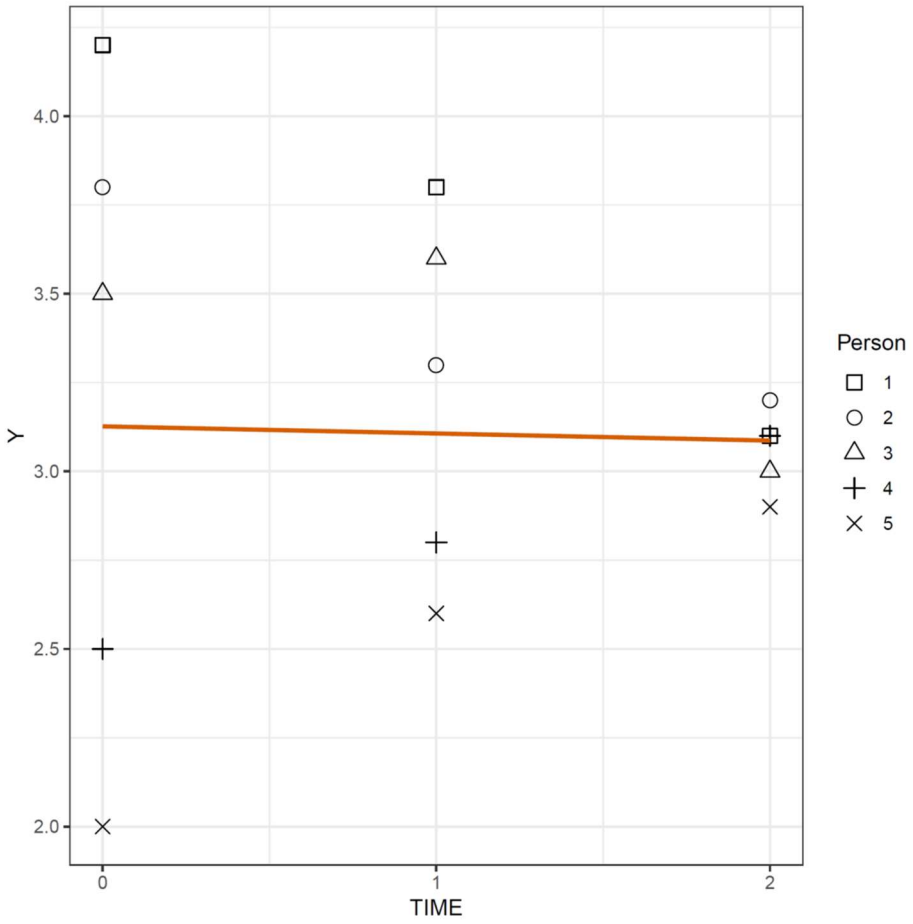

Figure 1. Example for a prototypical consensus emergence pattern.

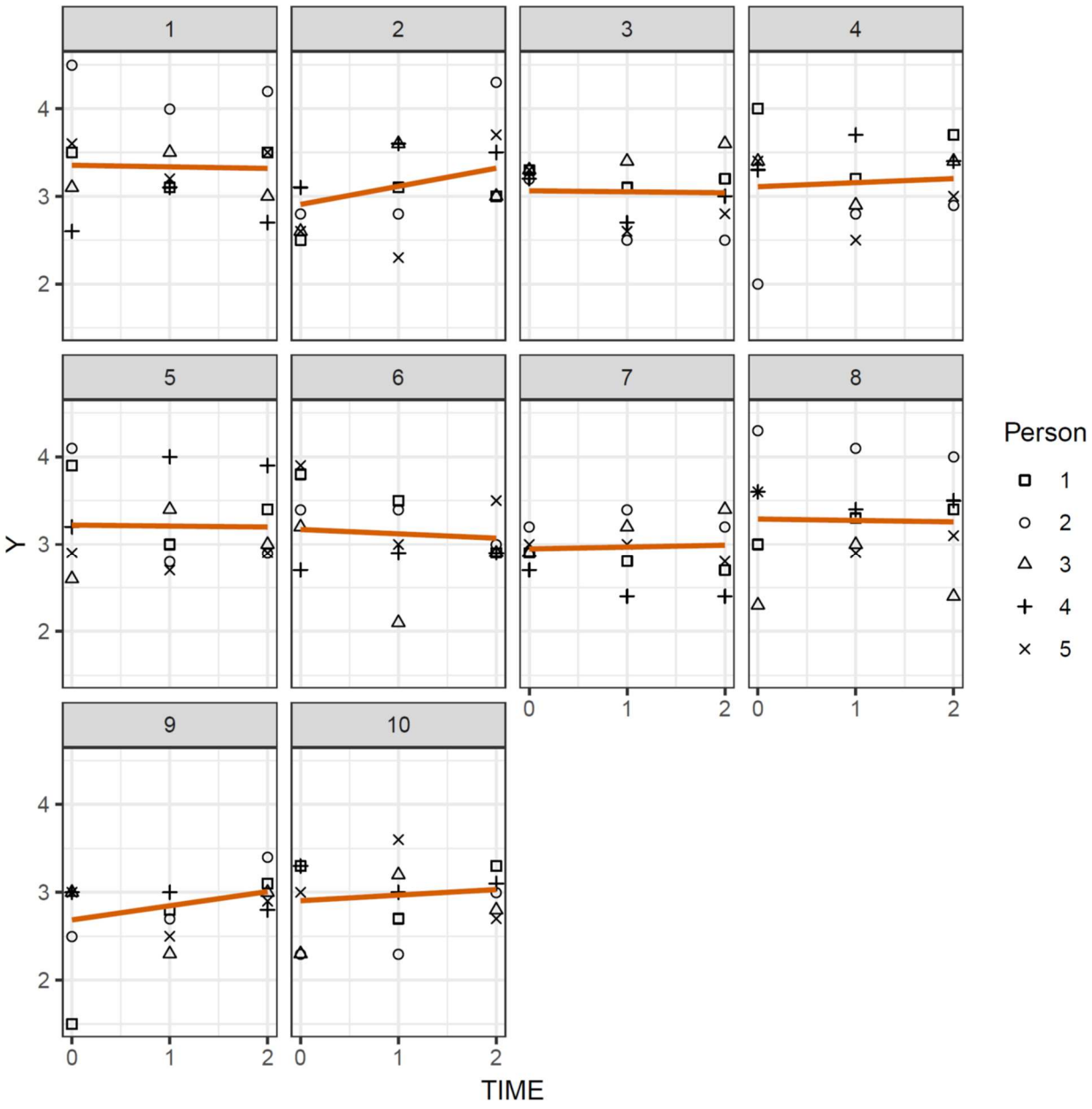

Figure 2. Plot of the example dataset in Table 1.

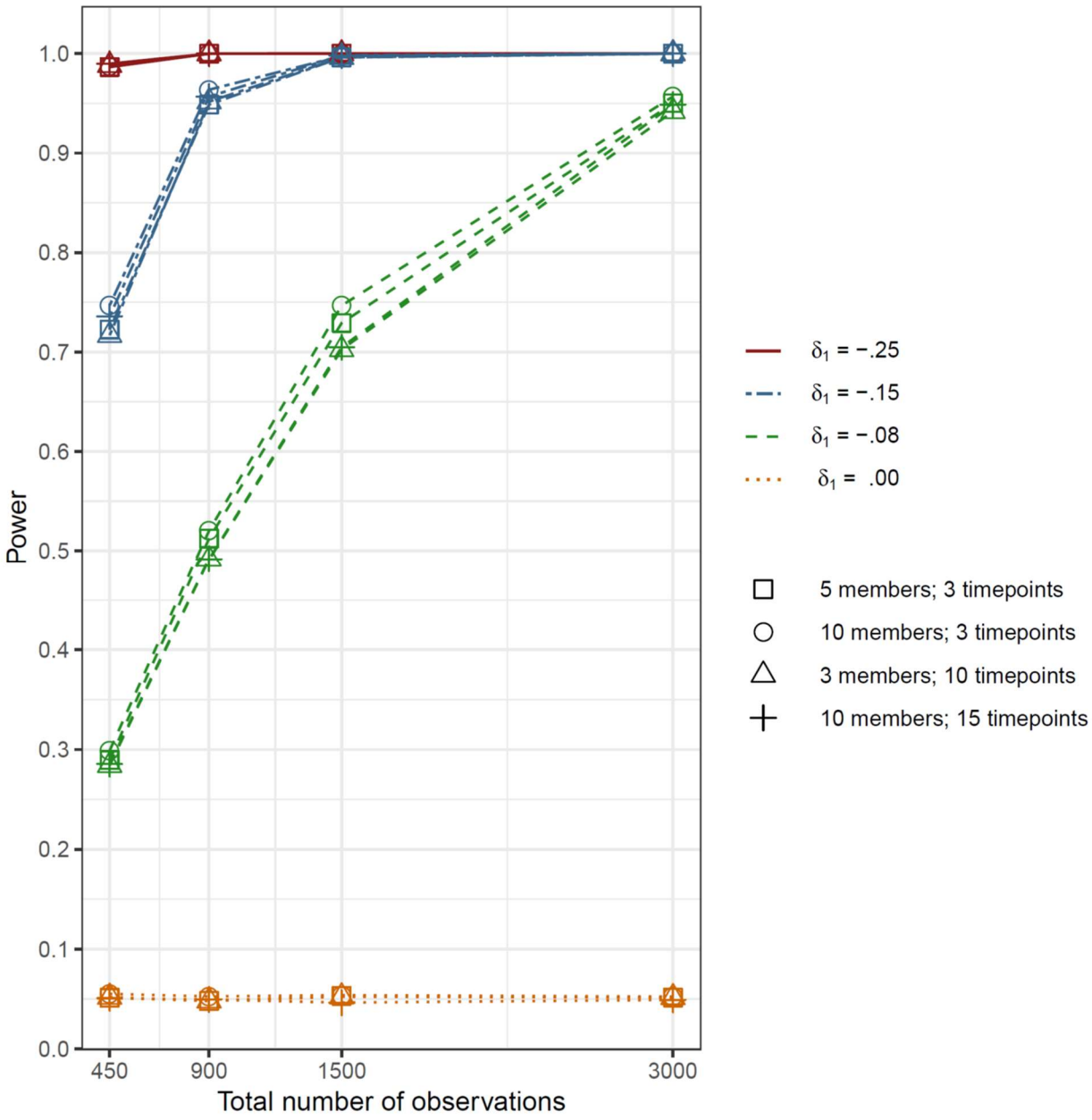

Figure 3. Power to detect a consensus emergence effect as a function of effect size, the total number of observations, and the distribution of the observations over groups and timepoints.

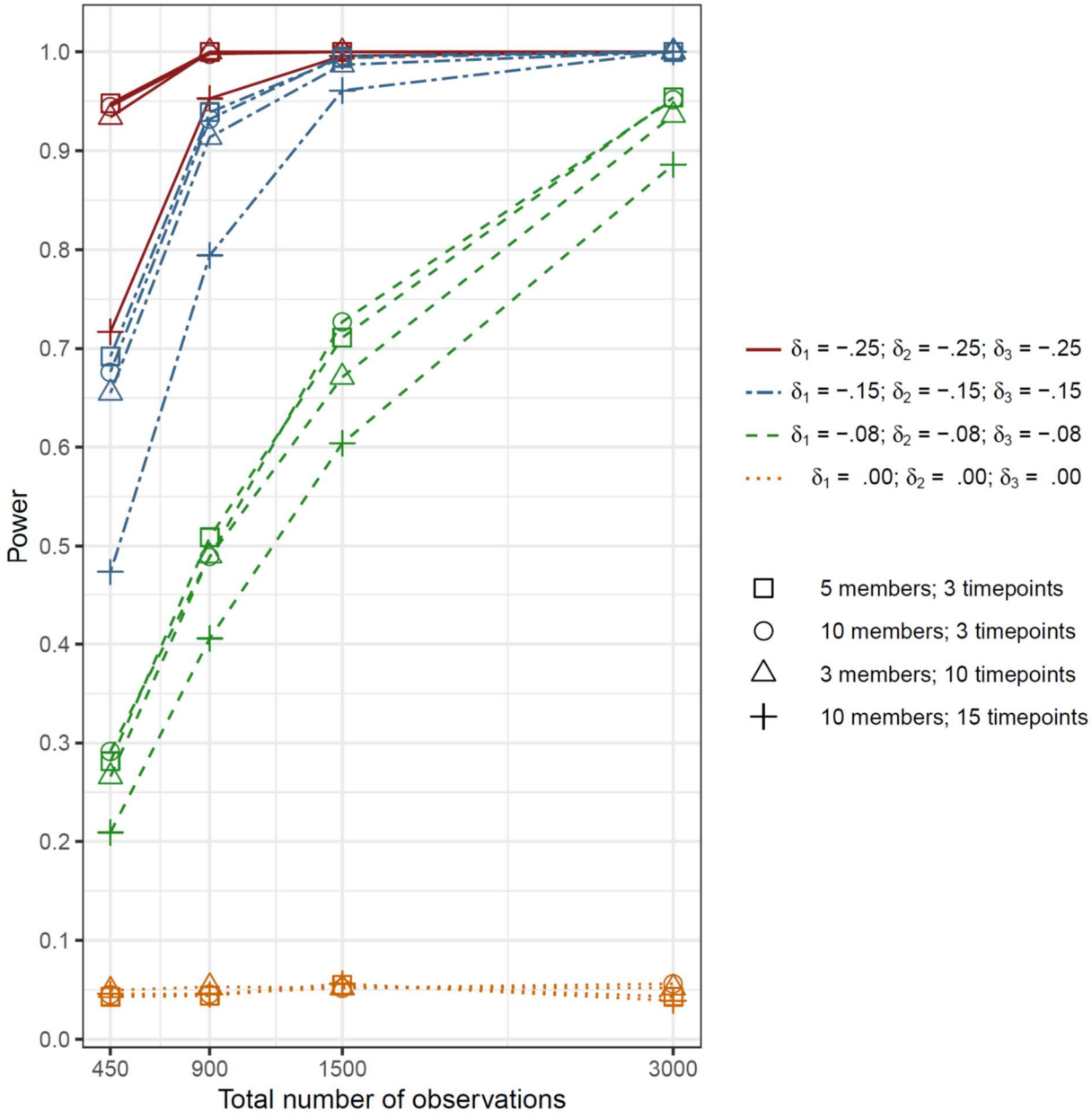

Figure 4. Power to detect a moderator of a consensus emergence effect as a function of effect size, the total number of observations, and the distribution of the observations over groups and timepoints.

Appendix A: Example R Data and R Code

```
edat<-expand.grid(time = 0:2,person = 1:5,group = 1:10)
edat$y <-c(3.5, 3.1, 3.5, 4.5, 4, 4.2, 3.1, 3.5,
3, 2.6, 3.1, 2.7, 3.6, 3.2, 3.5, 2.5, 3.1, 3, 2.8, 2.8, 4.3,
2.6, 3.6, 3, 3.1, 3.6, 3.5, 2.6, 2.3, 3.7, 3.3, 3.1, 3.2, 3.2,
2.5, 2.5, 3.3, 3.4, 3.6, 3.2, 2.7, 3, 3.3, 2.6, 2.8, 4, 3.2,
3.7, 2, 2.8, 2.9, 3.4, 2.9, 3.4, 3.3, 3.7, 3.4, 3.4, 2.5, 3,
3.9, 3, 3.4, 4.1, 2.8, 2.9, 2.6, 3.4, 3, 3.2, 4, 3.9, 2.9, 2.7,
2.9, 3.8, 3.5, 2.9, 3.4, 3.4, 3, 3.2, 2.1, 2.9, 2.7, 2.9, 2.9,
3.9, 3, 3.5, 2.9, 2.8, 2.7, 3.2, 3.4, 3.2, 2.9, 3.2, 3.4, 2.7,
2.4, 2.4, 3, 3, 2.8, 3, 3.3, 3.4, 4.3, 4.1, 4, 2.3, 3, 2.4, 3.6,
3.4, 3.5, 3.6, 2.9, 3.1, 1.5, 2.8, 3.1, 2.5, 2.7, 3.4, 3, 2.3,
3, 3, 3, 2.8, 3, 2.5, 2.9, 3.3, 2.7, 3.3, 2.3, 2.3, 3, 2.3, 3.2,
2.8, 3.3, 3, 3.1, 3, 3.6, 2.7)
edat$pred<-rep(c(0.2,2.0,1.0,-0.3,-1.0,-0.3,-0.2,0.1,0.1,0.4),each=15)

library(nlme)

m1<-lme(y ~ time, random = list(group=pdSymm(~time),
             person=pdIdent(~1)),data=edat)
m2<-update(m1,weights=varExp( form = ~ time))
anova(m1,m2)

m3<-lme(y ~ time*pred, random = list(group=pdSymm(~time),
             person=pdIdent(~1)),data=edat,
             weights=varComb(varExp( form = ~ time),varExp( form = ~ pred)))
m4<-update(m3,weights=varComb(varExp( form = ~ time),
                    varExp( form = ~ pred),varExp( form = ~ pred*time)))
anova(m3,m4)

nu00<-as.numeric(VarCorr(m2)[2,1])
nu11<-as.numeric(VarCorr(m2)[3,1])
nu01<-as.numeric(VarCorr(m2)[3,3])*(
             as.numeric(VarCorr(m2)[2,2])*as.numeric(VarCorr(m2)[3,2]))
tau<-as.numeric(VarCorr(m2)[5,1])
vsigma<-as.numeric(VarCorr(m2)[6,1])
delta1<-m2$modelStruct$varStruct
time<-0:2

# ICCEM
((nu00+2*nu01*time+nu11*time^2)/
(nu00+2*nu01*time+nu11*time^2+tau+vsigma*exp(2*delta1*time)))

r2lr<-function(mod1,mod0=NULL) {
       if(is.null(mod0)&is(mod1,"lm")) {
       mod0<-update(mod1,.~1)
       }
       if(is.null(mod0)&is(mod1,"merMod")) {
       mod0<-lm(update(formula(mod1),.~1),data=mod1@frame)
       }
       if(is.null(mod0)&is(mod1,"lme")) {
       mod0<-lm(update.formula(formula(mod1),.~1),data=mod1$data)
       }
       out<-1 - exp(-2/nobs(mod1) * (as.vector(logLik(mod1, REML = FALSE)) -

                            as.vector(logLik(mod0, REML = FALSE))))
       return(r2lr=out)
}

r2lr(m1)
r2lr(m2)
r2lr(m3)
```

```
r2lr(m4)
```

Appendix B: Power Simulation Code for the Consensus Emergence Model.

Note: Depending on the nature of the computer, running the full simulation may take up to several weeks. We recommend doing a test run with a small number of simulation runs before starting the full simulation

```
library(nlme)
# simulates a dataset with consensus emergence and fits a CEM to it
# function determines power of log-likelihood ratio test for delta1
simem <- function(l3n,l2n,l1n,
                                  gamma000,gamma100,
                                  nu00,nu11,nu01,tau,vsigma,
                                  delta1,tscale=NULL){
       if (is.null(tscale)) { tscale=l1n }
       dat=expand.grid(time = 0:(l1n-1)/((l1n-1)/(tscale-1)),
              member = 1:l2n,unit = 1:l3n)
       u <- MASS::mvrnorm(l3n, c(0,0), matrix( c(nu00,nu01,nu01,nu11), 2) )
       dat$u1<-u[,1][dat[,3]]
       dat$u2<-u[,2][dat[,3]]
       dat$r<-rep(rnorm(l2n*l3n,0,sd=sqrt(tau)),each=l1n)
       dat$e<-rnorm(l1n*l2n*l3n,0,sd=sqrt(vsigma*exp(2*delta1*dat$time)))
       dat$y<-(gamma000+gamma100*dat$time+dat$u1+dat$u2*dat$time+dat$r+dat$e)
       m1<-try(lme(y ~ time, random = list(unit=pdSymm(~time),
              member=pdIdent(~1)),data=dat,
              control=lmeControl(maxIter=15000,msMaxIter=15000)))
       m2<-try(update(m1,weights=varExp( form = ~ time)))
       if((inherits(m1, 'try-error')==F)&&(inherits(m2, 'try-error')==F)) {
              evsigma<-as.numeric(VarCorr(m2)[6,1])
              etau<-as.numeric(VarCorr(m2)[5,1])
              enu00<-as.numeric(VarCorr(m2)[2,1])
              enu11<-as.numeric(VarCorr(m2)[3,1])
              enu01<-as.numeric(VarCorr(m2)[3,3])*(
                     as.numeric(VarCorr(m2)[2,2])*as.numeric(VarCorr(m2)[3,2]))
              edelta1<-m2$modelStruct$varStruct
              out<-c(anova(m1,m2)[2,9],fixef(m2),enu00,enu11,enu01,
                                  etau,evsigma,edelta1)
       } else {     out<-rep(NA,9)      }
       return(out)
}

# illustrative use of simempred for a single situation
set.seed(123)
REPS=3 # test purposes only
#REPS=1000 # uncomment for actual simulation
system.time(simresults<-sapply(1:REPS, function(i,...) {
                    set.seed(123+i); simem(l3n=30,l2n=3,l1n=10,
                    gamma000=3,gamma100=0.01,
                    nu00=0.05,nu11=0.005,nu01=-0.005,tau=0.10,vsigma=.20,
                    delta1=-0.15,tscale=3)}))
mean(simresults[1,] < 0.05 ,na.rm=T)


# full simulation
sdat<-data.frame(
       l3n = rep(c(30,60,100,200,15,30,50,100,15,30,50,100,3,6,10,20),4),
```

```
       l2n = rep(rep(c(5,10,3,10),each=4),4),
       l1n = rep(rep(c(3,3,10,15),each=4),4),
       delta1 = rep(c(-0.25,-0.15,-0.08,0),each=16),
       tn=NA,power=NA,nran=NA,gamma000=NA,gamma100=NA,
                    nu00=NA,nu11=NA,nu01=NA,tau=NA,
                    vsigma=NA,edelta1=NA)
sdat$tn<-sdat$l3n*sdat$l2n*sdat$l1n

library(parallel)
no_cores <- detectCores() - 1  # get the number of cores
no_cores

runrows<-1:nrow(sdat) # which conditions
REPS = 10000 # simulation runs
path1<-"C:\\mydata\\test1.RData" # where to save the workspace

cl<-makeCluster(no_cores)
clusterExport(cl,ls())
clusterEvalQ(cl, library("nlme"))
system.time(
out<-parSapply(cl,runrows, function(i,...) {
       set.seed(123+i);
       simresults<-sapply(1:REPS, function(j,...) {
              simem(sdat[i,]$l3n,
              sdat[i,]$l2n,sdat[i,]$l1n,
              gamma000=3,gamma100=0.01,
              nu00=0.05,nu11=0.005,nu01=-0.005,tau=0.10,vsigma=.20,
              sdat[i,]$delta1,tscale=3)})
       return(c(mean(simresults[1,] < 0.05 ,na.rm=T),
              table(is.na(simresults[1,]))[1],
              rowMeans(simresults,na.rm=T)[2:9]))
       })
)
stopCluster(cl)
sdat1<-sdat
sdat1[runrows,6:15]<-t(out)
sdat1[runrows,]

saveRDS(path1)
```

Appendix C: Power Simulation Code for the Consensus Emergence Model With a Predictor

```
library(nlme)

# simulates a dataset with consensus emergence that is explained
# by a predictor and fits a CEM to it
# function determines power of logliklihood ratio test for delta3
simempred <- function(l3n,l2n,l1n,
                                  gamma000,gamma100,gamma010,gamma110,
                                  nu00,nu11,nu01,tau,vsigma,
                                  delta1,delta2,delta3,tscale=NULL){
       if (is.null(tscale)) { tscale=l1n }
       dat=expand.grid(time = 0:(l1n-1)/((l1n-1)/(tscale-1)),
              member = 1:l2n,unit = 1:l3n)
       u <- MASS::mvrnorm(l3n, c(0,0), matrix( c(nu00,nu01,nu01,nu11), 2) )
       dat$u1<-u[,1][dat[,3]]
       dat$u2<-u[,2][dat[,3]]
       dat$pred<-rep(rnorm(l3n),each=l1n*l2n)
       dat$r<-rep(rnorm(l2n*l3n,0,sd=sqrt(tau)),each=l1n)
       dat$e<-rnorm(l1n*l2n*l3n,0,sd=sqrt(vsigma*
                           exp(2*delta1*dat$time)*exp(2*delta2*dat$pred)*
                           exp(2*delta3*(-1)*dat$time*dat$pred)))
```

```
        dat$y<-(gamma000+gamma100*dat$time+gamma010*dat$pred+

        gamma110*dat$time*dat$pred+dat$u1+dat$u2*dat$time+dat$r+dat$e)
        m1<-try(lme(y ~ time*pred, random = list(unit=pdSymm(~time),
               member=pdIdent(~1)),data=dat,
               control=lmeControl(maxIter=15000,msMaxIter=15000),
               weights=varComb(varExp( form = ~ time),varExp( form = ~ pred))))
        m2<-try(update(m1,weights=varComb(varExp( form = ~ time),
                       varExp( form = ~ pred),varExp( form = ~ pred*time))))
        if((inherits(m1, 'try-error')==F)&&(inherits(m2, 'try-error')==F)) {
               evsigma<-as.numeric(VarCorr(m2)[6,1])
               etau<-as.numeric(VarCorr(m2)[5,1])
               enu00<-as.numeric(VarCorr(m2)[2,1])
               enu11<-as.numeric(VarCorr(m2)[3,1])
               enu01<-as.numeric(VarCorr(m2)[3,3])*(

        as.numeric(VarCorr(m2)[2,2])*as.numeric(VarCorr(m2)[3,2]))
               edelta1<-m2$modelStruct$varStruct$A
               edelta2<-m2$modelStruct$varStruct$B
               edelta3<-m2$modelStruct$varStruct$C
               out<-c(anova(m1,m2)[2,9],fixef(m2),enu00,enu11,enu01,
                                    etau,evsigma,edelta1,edelta2,edelta3)
        } else {      out<-rep(NA,13)      }
        return(out)
}

#illustrative use of simempred for a single situation
set.seed(321)
REPS=3 # test purposes only
#REPS=1000 # uncomment for actual simulation
system.time(simresults2<-sapply(1:REPS, function(i,...) {
                       set.seed(321+i); simempred(l3n=30,l2n=3,l1n=10,
                       gamma000=3,gamma100=0.01,gamma010=0.01,gamma110=0.01,
                       nu00=0.05,nu11=0.005,nu01=-0.005,tau=0.10,vsigma=.20,
                       delta1=-0.15,delta2=-0.15,delta3=-0.15,tscale=3)}))
mean(simresults2[1,] < 0.05 ,na.rm=T)

# full simulation
sdat<-data.frame(
        l3n = rep(c(30,60,100,200,15,30,50,100,15,30,50,100,3,6,10,20),4),
        l2n = rep(rep(c(5,10,3,10),each=4),4),
        l1n = rep(rep(c(3,3,10,15),each=4),4),
        delta1 = rep(c(-0.25,-0.15,-0.08,0),each=16),
        delta2 = rep(c(-0.25,-0.15,-0.08,0),each=16),
        delta3 = rep(c(-0.25,-0.15,-0.08,0),each=16),
        tn=NA,power=NA,nran=NA,gamma000=NA,gamma100=NA,gamma010=NA,gamma110=NA,
                       nu00=NA,nu11=NA,nu01=NA,tau=NA,
                       vsigma=NA,edelta1=NA,edelta2=NA,edelta3=NA)
sdat$tn<-sdat$l3n*sdat$l2n*sdat$l1n

library(parallel)
no_cores <- detectCores() - 1  # get the number of cores
no_cores

runrows<-1:nrow(sdat) # which conditions
REPS = 10000 # simulation runs
path2<-"C:\\mydata\\test2.RData" # where to save the workspace

cl<-makeCluster(no_cores)
clusterExport(cl,ls())
clusterEvalQ(cl, library("nlme"))
system.time(
out<-parSapply(cl,runrows, function(i,...) {
```

```
        set.seed(321+i);
        simresults<-sapply(1:REPS, function(j,...) {
               simempred(sdat[i,]$l3n,
                      sdat[i,]$l2n,sdat[i,]$l1n,
                      gamma000=3,gamma100=0.01,gamma010=0.01,gamma110=0.01,
                      nu00=0.05,nu11=0.005,nu01=-0.005,tau=0.10,vsigma=0.20,
                      sdat[i,]$delta1,sdat[i,]$delta2,sdat[i,]$delta3,
                      tscale=3)})
        return(c(mean(simresults[1,] < 0.05 ,na.rm=T),
               table(is.na(simresults[1,]))[1],
               rowMeans(simresults,na.rm=T)[2:13]))
        })
)
stopCluster(cl)
sdat2<-sdat
sdat2[runrows,8:21]<-t(out)
sdat2[runrows,]

saveRDS(path2)
```