J Bus Psychol (2018) 33:1–23
DOI 10.1007/s10869-017-9491-z

ORIGINAL PAPER

Back to Basics with Mixed-Effects Models: Nine Take-Away Points

Paul D. Bliese[1] · **Mark A. Maltarich**[1] · **Jonathan L. Hendricks**[1]

Published online: 5 April 2017

Abstract
Purpose Multilevel mixed effects models are widely used in organizational behavior and organizational psychology to test and advance theory. At times, however, the complexity of the models leads researchers to draw erroneous inferences or otherwise use the models in less than optimal ways. We present nine take-away points intended to enhance the theoretical precision and utility of the models.
Approach We demonstrate our points using two types of simulated data: one in which group membership is irrelevant, and the other in which relationships exist only because of group membership. We then demonstrate that the effects we observe in simulated data replicate in organizational data.
Findings Little that we address will be new to methodology experts; nonetheless, we draw together a variety of points that we believe will help advance both theory and analytic rigor in multilevel analyses.
Implications We make two points that run somewhat counter to conventional norms. First, we argue that mixed-effects models are appropriate even when ICC(1) values associated with the outcome data are small and non-significant. Second, we show that high ICC(2) values are not a prerequisite for detecting emergent multilevel relationships.
Originality/Value The article is designed to be a resource for researchers who are learning about and applying mixed-effects (i.e., multilevel) models.

Keywords Multilevel · Cross-level · Reliability · Centering · Emergence · ICC

✉ Paul D. Bliese
Paul.bliese@moore.sc.edu

Mark A. Maltarich
Mark.maltarich@moore.sc.edu

Jonathan L. Hendricks
Jonathan.Hendricks@grad.moore.sc.edu

[1] Darla Moore School of Business, University of South Carolina, Columbia, SC, USA

Mixed-effect or multilevel models such as HLM (Raudenbush and Bryk 2002), PROC Mixed in SAS (Littell et al. 2006; Wolfinger 1997), and lme in R (Pinheiro and Bates 2000) are routinely used to analyze data in organizational behavior and organizational psychology. Researchers recognize that nested data are often non-independent such that responses on the dependent variable from members of the same group are more similar than would be expected by chance (Bliese 2000). The idea that mixed-effects models can account for non-independence is well documented; however, options surrounding model specification and interpreting parameter estimates from mixed-effects models are not always straightforward. As such, researchers occasionally underutilize mixed-effects models, engage in inappropriate model building, or misinterpret model parameters.

Our goals are to (a) clarify situations where mixed-effects models may be useful and (b) explain the interpretation of results in common variants of the mixed-effects model. Much of what we cover will be known to methodologists; nonetheless, we believe that going back to basic ideas can help researchers more effectively use these methods to test and advance theory. We also emphasize that two of the points we raise (using mixed-effects models when levels of non-independence are minimal and detecting emergent effects when group-mean reliability is low) run counter to conventional norms, so we encourage authors, editors, and reviewers to reconsider these two points in particular.

On certain occasions we will draw attention to specific articles and refer to the lead author's experience as an Associate Editor over the last 7 years. We want to stress that our goal is not to be critical of other authors. Rather, we recognize that the field evolves and the complexity of these methods makes it easy to make inferences that might later be interpreted differently—indeed, we identify one example where the lead author and his colleagues specified hypotheses in ways that would now be stated otherwise. At the same time, we do discuss some specific articles because we want to demonstrate that the issues we identify are relatively common as a basis for advocating improvements.

Understanding the Extremes

We take what we believe to be a novel approach to describing the models. Specifically, we contrast ordinary least squares (OLS) with mixed-effect regression models on data from two ends of a continuum. At one end of the continuum, all data are independent. These data represent a situation where a common measure of non-independence, the Intraclass Correlation Coefficient (1) or ICC(1) (see Bliese 2000), is effectively zero. In this situation, there is no similarity in responses among members of the same group, so group membership is irrelevant to the data analysis. Randomly assigning group membership to individuals produces this form of data. Applying a mixed-effects model to data at this end of the continuum is rarely done; however, we show it is nonetheless informative to contrast mixed-effects and OLS regression under these conditions.

At the other end of our continuum, we consider the situation where the relationship between two lower-level variables (predictor and outcome) is only a function of group membership. At this part of the continuum, the ICC(1) is greater than zero and is both statistically and practically significant. In this situation, individual responses represent imperfect measures of a true group construct. Pure real-world examples at this end of the continuum are rare; however, later we describe in detail a conceptual example where individuals are randomly assigned to rooms that vary in terms of physical temperature and are asked to estimate the temperature.[1] What is important for our arguments, though, is that this end of the continuum provides the logical contrast to the situation where group-level properties are zero.

In practice, organizational data generally fall between these two extremes. The magnitude of raw correlations between constructs is frequently a function of both individual-level and group-level factors. In addition, the levels of non-independence associated with group membership [i.e., the ICC(1)] for common constructs such as well-being are often lower than .05 (Bliese 2006; Murray and Short 1995). Therefore, we augment our analyses of simulated data from the two ends of the continuum by illustrating the majority of our take-away points in a large organizational dataset where the dependent variable has an ICC(1) value less than 0.05 and the mechanisms underlying the observed relationships were not predetermined in a simulation. In so doing, we show that the points we identify in simulated data apply to prototypical organizational data. We also use the organizational data to illustrate two points related to testing cross-level interactions—points that can be clearly made without relying on simulated data.

Simulation 1: Independent Data

In this section, we examine simulated data at the end of the continuum where data and relationships between variables are independent of group membership. Note that here and throughout, we use the terms "individual-level," "level-1," and "lower-level" to refer to data at the lowest-level of nesting (e.g., an individual in a group) and we use the terms "group-level," "level-2," or "higher-level" to refer to the higher level of nesting (characteristics of the group that would be consistent for all members of the same group). We also refer only to two-level models although many ideas we discuss would generalize to more complex data structures. In addition, the appendix provides R code for all data used in the examples.

The basic idea behind creating data where relationships are independent of group membership is to create two vectors of correlated numbers (X and Y) and randomly assign group membership to the (X,Y) pairs. Because multilevel analyses often involve aggregate variables (level-2 variables), the simulation code also calculates the group-mean of X (X.G) and assigns the group-mean back to each "individual." The first 15 rows of the data generated by the simulation code are presented in Table 1. These data have no group-level properties. Using a mixed-effects model (lme), the ICC(1) for our *Y* variable is 4.3×10^{-09}, effectively zero.[2]

Mixed-Effects Models on Independent Data

Our first take-away states that "misapplying" mixed-effects models to data that could be analyzed in simple ordinary least squares (OLS) regression has little (if any) downside. In the case where the ICC(1) for the dependent variable is zero, a

[1] We are indebted to the late Larry James who (to our knowledge) was the originator of this example and used it when discussing these ideas in conferences.

[2] There are two ways to calculate ICC(1)—using ANOVA or mixed-effects models (Bartko 1976; Bliese 2000). Using the ANOVA method here would produce a slightly negative ICC(1), which is sometimes truncated to zero in reporting. The differences between the two methods are generally not substantial, but the decision point is worth noting.

Table 1 First 15 rows of simulated data

G.ID	Y	X	X.G
1	−0.356	−0.253	−0.181
1	−0.422	−1.108	−0.181
1	−0.475	0.123	−0.181
1	−0.075	−0.209	−0.181
1	0.040	−0.565	−0.181
1	−0.043	0.318	−0.181
1	1.847	0.923	−0.181
1	0.555	−0.504	−0.181
1	1.654	0.784	−0.181
1	−2.321	−1.314	−0.181
2	−0.828	−0.661	−0.206
2	0.393	1.917	−0.206
2	0.373	−0.808	−0.206
2	−2.492	−2.259	−0.206
2	−1.405	1.587	−0.206

G.ID group identifier, *X.G* group mean of X

random intercept mixed-effects model returns parameter estimates and standard errors that are virtually identical to regression results. Admittedly, the results returned by the mixed-effects models are more computationally intensive and are based on restricted maximum likelihood estimation algorithms, but assuming the mixed-effects model converges, it will return results that mirror those from OLS regression.

Take-away point 1: Mixed-effects models can be used with nested data, even if no group-level effects are expected or evident.

We elaborate using the simulated data. The upper part of Table 2 presents the results from regressing Y on X and X.G in OLS regression. Notice that X is significantly related to Y, but X.G (group means of X assigned back to individuals) is not. These results are expected. It would be surprising to find an emergent X.G effect in data with no group-level properties. As we discuss in several points in the article, an emergent effect (aka a contextual or incremental effect—see Hofmann and Gavin 1998) is formally a test of whether the strength of the relationship (e.g., slope) between individual-level variables differs from the strength of the relationship between group means of the same variables (Alwin 1976; Firebaugh 1980).

The lower part of Table 2 provides results from a mixed-effects model with a random intercept. There are two differences between the models. First, the mixed-effects model in lme in R provides degrees of freedom (DF) identifying X.G as a group-level variable (which we also return to in more detail later). Second, the *p* value associated with X.G is 0.266 instead of 0.263 which occurs because the *p* value is based on a *t* value with 98 DF in the lower model instead of being based on 997 DF in the top model. Notice, however, that all parameter estimates and standard errors are identical. Indeed, while Table 2 shows only three decimal places, the parameter estimates and standard errors are identical through the first eight decimal places (the default in R).[3]

Table 2 Results of regressing Y on X and X.G in OLS (upper panel) and mixed-effects random intercept model (lower panel) in simulated data with near-zero ICC(1)

	Estimate	Std. error	*DF*	*t* value	Pr(>|t|)
(Intercept)	0.062	0.030	–	2.035	0.042
X	0.302	0.032	–	9.552	0.000
X.G	−0.110	0.098	–	−1.119	0.263
(Intercept)	0.062	0.030	899	2.035	0.042
X	0.302	0.032	899	9.552	0.000
X.G	−0.110	0.098	98	−1.119	0.266

Note: $N = 1000$ individuals in 100 groups

A slight modification in the simulation allows us to illustrate the second point:

Take-away point 2: Failing to use mixed effects to model nested data can increase the risk of type I error (for group-level effects) and type II error (for lower-level effects).

If we re-run the simulation with a slightly different random seed of 125,321, we observe an ICC(1) estimate of 0.013 (1.3% of the variance in Y is attributable to group membership). From our simulation procedure, we know the ICC(1) value is just an artifact. An *F*-test of the ANOVA model indicates that an ICC(1) of 0.013 is not statistically significant ($p = .194$), so these data still represent what would be widely considered independent data with no significant group-level effects. In these data, however, the OLS regression model and mixed-effects model now differ in predictable (but occasionally misunderstood) ways.

Table 3 provides OLS and random intercept mixed-effects model results. The parameter estimates are identical, but even the small, non-significant, ICC(1) value of 0.013 impacts the

Table 3 Results of regressing Y on X and X.G in OLS (upper panel) and mixed-effects random intercept model (lower panel) in simulated data with small positive ICC(1)

	Estimate	Std. error	*DF*	*t* value	Pr(>|t|)
(Intercept)	0.015	0.031	–	0.482	0.630
X	0.283	0.033	–	8.605	0.000
X.G	0.095	0.108	–	0.877	0.381
(Intercept)	0.015	0.032	899	0.463	0.644
X	0.283	0.033	899	8.645	0.000
X.G	0.095	0.112	98	0.845	0.400

Note: $N = 1000$ individuals in 100 groups

[3] Econometricians have noted that random effect models can yield level 1 effects that are biased because they "inherit" group-level effects. Including group means as a predictor in random effect models removes this bias and produces coefficients for x that are comparable to fixed-effects models (see Raudenbush 2009).

standard errors. As detailed in Bliese and Hanges (2004) and Pinheiro and Bates (2000), the standard errors behave differently for the two classes of predictors. Specifically, the mixed-effects standard error for X (lower panel results) is 0.03269 and slightly smaller than the OLS standard error for X (upper panel) of 0.03285 (both round to 0.033 in Table 3). In contrast, in the mixed-effects model, the standard error for X.G (lower panel) is slightly larger than the standard error for X.G in the OLS model (upper panel). Stated otherwise, ignoring the non-independence in the data, the OLS model is slightly conservative regarding significance tests involving X (level-1 variables) leading to type II error, but it is slightly liberal regarding significant tests involving X.G (level-2 variables) leading to type I error. Large ICC(1) values magnify these effects on the standard errors (see Bliese and Hanges 2004).

It is common to see warnings about how ignoring non-independence in OLS models can lead to overly liberal tests of significance. For instance, Aguinis and Molina-Azorin (2015) write that "Specifically, when using techniques that rely on the independence assumption, as is the case for ordinary least squares regression, the resulting standard errors will be downwardly biased, resulting in Type I statistical errors" (p. 354). This warning is well founded; however, the warning really applies to higher-level variables (X.G in our example). Ignoring non-independence also leads to overly conservative tests of lower-level variables.

Our example with the small ICC(1) value also reinforces the importance of take-away point number 1, which is that there is no real downside to estimating a random effect model. Because even small ICC(1) values can bias standard errors, one is almost always better off using mixed-effects models for data that come from groups where non-independence might be present. At a practical level, authors can use mixed-effects models when they have low ICC(1) values and editors and reviewers should not hesitate to request mixed-effects models if data are nested even if the ICC(1) values are small. Even in cases where multiple levels of nesting exist (e.g., soldiers in platoons in companies in battalions), it is relatively simple to account for non-independence at each level to help ensure that standard errors are correct (for instance, Lang et al. 2007 reported a five-level nested random intercept model to control for levels of non-independence in a military sample).

It is worth commenting on the fact that McNeish and colleagues have raised issues about the "unnecessary ubiquity" of mixed-effect models (McNeish et al. 2017). Importantly, however, the key argument made by McNeish and colleagues revolves around the fact that there are a variety of methods to adjust for non-dependence including generalized estimation equations (gee), the use of robust standard errors, and adding fixed-effects for group membership (e.g., Verbeek 2008). In some cases, models other than mixed-effects may be preferred, but in our mind the bigger issue is the need to adjust for non-independence using some reasonable method, and in most cases, mixed-effects models represent a reasonable method.

As a final practical observation related to points 1 and 2, we note that authors often write text implying that the use of mixed-effects models must be "justified" by large ICC(1) values (Li et al. 2016; Liu et al. 2012; Miron-Spektor et al. 2011; Wang and Howell 2010). In contrast, we suggest that authors should have to justify their decision *not* to use mixed-effects models (or other approaches that can control for non-independence) by showing that ICC(1) values are so trivial as to be considered zero. Ultimately, though, it is often difficult to make a compelling case that the ICC(1) so small as to be ignorable, so a preferable strategy would be to routinely default to mixed-effects models or equivalent if data have been collected from groups.

Group-Mean Centering on Independent Data

The next variation on our simulation lets us illustrate issues surrounding centering level-1 variables and is summarized in take-away point number 3:

Take-away point 3: Group-mean centering of a level-1 variable fundamentally changes the interpretation of the level-2 parameter estimate for the analogue of the same variable. In group-mean centered models, level-2 parameter estimates represent overall group effects; in raw or grand-mean models level-2, parameter estimates represent differences in slopes between individual-level and group-level relationships.

One of the common decision points authors deal with when working with mixed-effects data is whether to center level-1 variables (Hofmann and Gavin 1998). Three centering options are possible: raw (no centering), grand-mean centering (subtracting a variable's overall mean from each observation), and group-mean centering (subtracting a group's mean on the variable from each observation). Raw data and grand-mean centering produce equivalent models, with a slight technical advantage going to grand-mean centering because it reduces the correlation between slopes and intercepts (Hofmann and Gavin 1998; Kreft and De Leeuw 1998). Group-mean centering, in contrast, can substantively change model interpretation.

Texts differ on whether group-mean centering is to be encouraged or not. For instance, Raudenbush and Bryk (2002) favor group-mean centering while texts such as Snijders and Bosker (1999) are less favorably oriented. Good cases can be made both for and against group-mean centering. There is, however, one clear case where group-mean centering should be used to verify results. The specific case involves analyses examining cross-level interactions (i.e., where a group-level variable moderates the relationship between a level-1 predictor and the level-1 outcome) which we discuss in the examples utilizing organizational data.

Table 4 Results of regressing Y on W.X and X.G in OLS (upper panel) and mixed-effects random intercept model (lower panel) in simulated data with small positive ICC(1)

	Estimate	Std. error	*DF*	*t* value	Pr(>|t|)
(Intercept)	0.015	0.031	–	0.482	0.630
W.X	0.283	0.033	–	8.605	0.000
X.G	0.377	0.103	––	3.666	0.000
(Intercept)	0.015	0.032	899	0.463	0.644
W.X	0.283	0.033	899	8.645	0.000
X.G	0.377	0.107	98	3.519	0.001

Note: N = 1000 individuals in 100 groups

While group-mean centering has some advantages when modeling cross-level interactions, defaulting to group-mean centered variables can be risky particularly when cross-level interactions are *not* of interest and the level-2 analogue of the same variable is included in the model. In Table 4, we have substituted X with a group-mean centered value of X, W.X (for within-group X). Notice that the OLS and mixed effects parameter estimates for W.X are the same (0.283) as parameter estimates for X in Table 3. That is, all models return the same estimate for the individual effect associated with X. Importantly, however, the models in Table 4 show that the level-2 variable, X.G, is significant. The effect associated with X.G *appears* to be an emergent effect where the relationship between group means *appears* to be significantly different than relationships observed at lower-levels (Alwin and Hauser 1975; Firebaugh 1980; Hofmann and Gavin 1998).

If we interpret the results for X.G in Table 4 as emergent effects, we should be troubled by the fact that the data simulation process that produced these data have no meaningful group-level properties. Therefore, identifying a significant emergent effect associated with X.G seems unlikely. Indeed, Table 4 illustrates a well-known phenomenon. When using group-mean centering, the level-2 effect "inherits" level-1 effects and represents the total level-2 effect. Here, the value of .377 represents (with rounding error) the effect of .283 from the individual-level effect plus a (statistically nonsignificant) group-mean specific effect of 0.095 (X. G in Table 3). Unfortunately, while this effect is well known, the lead author (in his role as an Associate Editor) has seen several occasions where users of mixed-effects models have used group-mean centering and have incorrectly interpreted level-2 effects as emergent effects.

In short, group-mean centering should be used with care, particularly when including group-level analogues of level-1 variables. When level-1 variables are group-mean centered, one can mistakenly conclude the level-2 group-mean analogue of the lower-level variable represents a test of an emergent effect when it actually represents a test of the total effect. A logical extension of this point is that if researchers are interested in testing whether the level-2 relationship is significant (not whether it is *significantly different* from the level-1 relationship), group-mean centering is logical. Occasionally, authors have proposed that both individual and group-level relationships will be significant and have incorrectly interpreted level-2 parameter estimates from raw or grand-mean centered variables to test the hypothesis (e.g., the test of hypothesis 1 in Liao and Chuang 2007) suggesting that it is relatively easy to make this error and that editors and reviewers may miss this point.

Before concluding the section on centering, we add one additional take-away point that is more theoretically oriented than methodologically oriented regarding centering:

Take-away point 4: Group-mean centering changes the conceptual meaning of the level-1 construct such that the term now reflects relative position in a group rather than absolute values.

Obviously, a best practice in analytics is to align (a) theory and hypotheses with (b) specific analyses being tested. When using group-mean centering, one is testing whether an individual's position relative to his or her group is important and so hypotheses should acknowledge the group referent. As an example, there is an important conceptual difference between hypothesizing that "People high in X tend to be high on Y," (raw or grand-mean centered) and "people who are higher than their group on X also tend to be higher on Y than their group" (group-mean centered X).

In our example with the simulated data, the difference between group-mean centered and raw variables is trivial. A model regressing Y on only the group-mean centered variable (omitting X.G) returns an estimate of 0.283[4] whereas a model regression Y on X without X.G in the model returns an estimate of 0.291. These results show that in a case where group effects are essentially random error, the absolute relationship between X and Y is fundamentally the same as the relationship adjusted for group differences. In practice, though, one will often observe that results differ depending upon whether one uses raw and group-mean centered variables. Equally important, precision in terms of language helps align theory and analytics.

Interpreting Level-2 Coefficients

Point number five addresses interpreting models where the level-1 outcome is regressed on a level-2 predictor such as X.G.

[4] The estimate of 0.283 is identical to the values reported in Tables 3 and 4. Table 3 therefore shows that one can remove any group effects from X by group-mean centering or by including the group mean of X into the model as noted in footnote 3.

Table 5 Results of OLS group-mean regression of Y.G on X.G in simulated data with small positive ICC(1)

	Estimate	Std. error	*t* value	Pr(>|t|)
(Intercept)	0.015	0.032	0.463	0.645
X.G	0.377	0.107	3.519	0.001

Note: $N = 100$ groups

Take-away point 5: In mixed-effects models, higher-level variables predict the group means of the lower-level outcome.

Changes to the simulated data let us demonstrate point five. By calculating group means for both X and Y (X.G and Y.G), we can create a dataset that contains 100 (X.G, Y.G) pairs instead of 1000 (X, Y) pairs. Table 5 provides the results from running a simple OLS regression on the group means. Notice that the X.G results from Table 5 match the X.G results from the lower part of Table 4 indicating the group-level coefficient in the mixed-effect model is predicting the group-mean of Y from the group-mean of X. The more appropriate standard errors from the mixed-effects model match those in the group-level OLS regression.

The idea that level-2 variables predict group means of Y can be extended to cases where the predictor has no level-1 analogue. If we modify the simulation to add a constant random number (Z.G) to each group member, Z.G now represents a group-level variable that differs across groups, but is the same within groups. As before, these data have no real group-level properties (Y has the same ICC(1) value of 0.013).

Not surprisingly, adding Z.G (a random number) as a group-level variable provides little predictive power. Once again, however, if we estimate (a) a random intercept, mixed-effects model regressing Y on Z.G and (b) an OLS model of the 100 group means regressing Y.G on Z.G, we get the same parameter estimates (0.01746231) and very similar standard errors (0.03664744, 0.03664742) in the two models.

The fact that group-level variables predict group means of the outcomes in mixed-effects models is known among methodologists (LoPilato and Vandenberg 2015; Preacher et al. 2010; Raudenbush and Bryk 2002). Indeed, Raudenbush and Bryk (2002) specifically refer to models with group-level predictors and individual-level outcomes as "means as outcomes" models to emphasize that the dependent variable has functionally been transformed to a group mean. Unfortunately, this aspect of mixed-effects models is often misrepresented when discussing theoretical foundations of multilevel research. That is, it is fairly common to see theoretical models built on the premise that group-level variables will predict individual outcomes (including some from the lead author), but it is more accurate to state that group-level variables predict group means of individual outcomes.

Simulation 2: Pure Group Data

We now move to the other end of the continuum to consider cases where relationships between X and Y exist only because the group means for X and the group means for Y have a relationship. The appendix provides code creating 100 groups of size 10 with two variables (X and Y). Both X and Y were generated to have an ICC(1) value of 0.15. In the simulation, the underlying population correlation between the group mean for X and the group mean for Y is 1; however, the observed correlation is a function of group-mean reliability (which we discuss).

As mentioned earlier, one way to conceptually think about these data would be to imagine that we have 100 physical locations that vary in terms of temperature from cold to warm. Each location is randomly assigned 10 group members and each group member estimates the temperature (Y). On the next day, we obtain another set of 1000 participants, randomly assign these new participants in groups of 10 to the 100 physical locations (temperatures unchanged) and each group member estimates the temperature (X).

To simulate this type of design, we randomly assign values around a true group score. Row 1 has two ratings (X and Y), but conceptually these ratings are from different people on different days and the only connection is that they are rating the same group-level target (room temperature). Indeed, with these data one could randomly sort either the X or Y variable (or both) within groups so that rows within groups had different X,Y pairs and the results would be functionally unchanged.

The covariance theorem (Dansereau et al. 1984; Robinson 1950) provides a useful way to decompose the raw correlation into its within-group and between-group components. Using the theorem, we observe a raw correlation of 0.147, a between-group (i.e., group-mean) correlation of 0.610, and a within-group (i.e., group-mean centered) correlation of 0.003. Stated otherwise, the observed correlation between X and Y at the raw individual level is strictly a function of group-level processes.[5] The ICC(1) for what we arbitrarily designate as the Y variable in the simulation is 0.153. The differences here accentuate a risk of ignoring non-independence altogether because regressing Y on X would yield a coefficient based on the raw correlation of 0.147 that could be misinterpreted as an individual-level effect.

[5] The appendix provides code to create an X2 variable that randomly sorts the X variable on a group-by-group basis and then re-estimates the components of the covariance theorem. As anticipated, aligning Y variables to different X responses within a group has no meaningful impact on the results.

Table 6 Results of regressing Y on X and X.G in OLS (upper panel) and mixed-effects random intercept model (lower panel) in simulated data with substantial positive ICC(1)

	Estimate	Std. error	*DF*	*t* value	Pr(>|t|)
(Intercept)	0.050	0.033	–	1.519	0.129
X	0.003	0.036	–	0.092	0.927
X.G	0.621	0.073	–	8.500	0.000
(Intercept)	0.050	0.043	899	1.182	0.237
X	0.003	0.034	899	0.096	0.924
X.G	0.621	0.089	98	6.990	0.000

Note: N = 1000 individuals in 100 groups

Emergent Effects with Mixed-Effects Models

Table 6 provides the results of an OLS model and a mixed-effects random intercept model. Notice that parameter estimates between the two models are identical, but the standard errors are different. The form of the standard errors is similar to what we observed previously in the data with the weak ICC(1) of 0.013; however, in Table 6, the effects are more dramatic, particularly for the group-level variable (X.G). The under-estimated standard error of 0.073 in the OLS model produces an inflated *t* value of 8.50. In contrast, the larger (but correct) standard error in the mixed-effects model of 0.089 produces a *t* value of 6.99. Notice that standard errors for the level-1 variable (X) behave in the opposite fashion (slightly smaller in the mixed-effects model) reducing type II errors in the mixed-effects model, consistent with take-away point 2.

The mixed-effects model provides evidence of the emergent effect. The results indicate that the relationship between group means differs from the relationship between individual variables. Recall, an emergent effect is formally a test of whether the strength of the relationship between two variables differs across levels. In Table 6, we interpret X.G as being 0.621 larger than the individual slope of .003 so the total slope associated with the group means is 0.621 + .003 or 0.624. Interested readers can confirm that the total relationship is 0.624 by (a) using group-mean centered variables at level 1 (see appendix), (b) estimating a mixed-effects model regressing Y on X.G alone, or (c) calculating group means for X and Y and estimating an OLS model on the group means (X.G and Y.G).

Having now demonstrated an emergent effect, we can introduce take-away point number 6 which, like point 4, is more conceptual than methodological.

Take-away point 6. In mixed-effects models, emergent effects identifying different relationships between level-1 variables and their group mean analogs likely represent important changes in the meaning of constructs across levels.

Emergent effects across levels are interesting. For both practical and theoretical reasons, researchers have considered why relationships involving group means might differ from relationships involving lower-level variables (Barrick et al. 1998; Luciano et al. 2014). We discuss several examples. In an early article in this genre, Mathieu and Kohler (1990) examined absenteeism climate among transit operators working in five different garages. Their basic hypothesis was that the link between previous absenteeism (past 6 months) and subsequent time lost (current 6 months) might differ when examined at the garage level relative to the individual level, and they found evidence of an emergent effect they attributed to an "absence culture."

Table 2 in Mathieu and Kohler reported a significant level-2 garage effect of 0.16, and since Mathieu and Kohler were not using group-mean centered variables, the coefficient of 0.16 indicates that the total garage-level slope of 0.29 (obtained by adding 0.16 to 0.13) was stronger than the individual-level slope of 0.13. Note that in 1990 the test reported in Mathieu and Kohler involved an OLS regression-based test of garage-level absence, so the standard error was likely downwardly biased, but the idea is interesting both practically and theoretically, and at the time the methods were considered appropriate.

As another example, emergent effects expressed as correlations were identified over two decades ago involving ratings of work hours and ratings of well-being among US Army companies (Bliese and Halverson 1996). Results suggested that individual work hours were only weakly related to individual well-being (−.16 raw, −.11 group-mean centered); however, average work hours in an army company were strongly related to average well-being (−.71). We analyze these data in more detail later.

More recently, interest in emergent effects has appeared to wane; however, occasionally researchers continue to explore these ideas. For instance, González-Morales et al. (2012) investigated emergent effects associated with collective cynicism and individual cynicism. In the article, the hypotheses were stated in ways that contravene take-away point number five. More specifically, hypothesis 2.2 stated that "Perceived collective cynicism at Time 1 is a significant predictor of *individual* cynicism at Time 2 over and above previous levels of individual cynicism, job demands, and resources indicators at Time 1" (italics added). Recall from take-away point number five that level-2 constructs (collective cynicism at time 1) do not actually predict level-1 outcomes (*individual* cynicism at time 2). As such, a better alignment between hypotheses and analytics would have stated "perceived collective cynicism at Time 1 will predict *average group* cynicism at Time 2." We point out how the hypotheses were phrased not to be critical, but simply to acknowledge that the fields' understanding of how to align theory with methods evolves over time. The key findings related to emergence (discussed below) hold regardless of how the hypotheses were phrased.

The models estimated by Gonzáles-Morales et al. suggested that time 1 measures of collective cynicism (where teachers rated cynicism among their peers) had emergent effects on groups' ratings of time 2 cynicism. That is, the authors found

that the relationship between time 1 collective cynicism and time 2 cynicism was stronger at the group-level than at the individual level. From a theoretical perspective, emergent effects often suggest that constructs have undergone a shift in meaning (Bliese 2000; Bliese et al. 2007; Firebaugh 1978; Morgeson and Hofmann 1999). In their discussion, Gonzáles-Morales et al. provide several explanations for what aggregate measures of collective cynicism might reflect, including a shared lack of resources and negative emotional contagion.

In the end, Gonzáles-Morales et al. may not be able to definitively state what the collective measure reflects. The aggregate construct likely represents some type of fuzzy composition process (Bliese 2000) in which meaning across levels differs but also remains conceptually linked to the construct of cynicism. Regardless, the results suggest that ratings of collective cynicism reflect some form of unproductive work climate that is temporally stable over 6 to 7 months. As other examples of shift in meaning across levels, one can make convincing cases that collective ratings of work hours reflect externally mandated work requirements (few groups as a whole would work 14 h a day unless externally required) and that collective absenteeism reflects norms among groups about the appropriateness of missing work (Mathieu and Kohler 1990).

It is also possible that the shift in meaning could be tied to an increase in reliability across levels (see also Ostroff 1993). Interestingly, increased reliability may also reflect a subtle form of shift in meaning. At one extreme, a completely unreliable measure captures nothing shared or reproducible among group members and hence no attributes of a shared construct. At this extreme one can think of the group-level measure as error. At the other extreme, a highly reliable measure captures something shared and reproducible in terms of group member ratings. We contend that a change from "random error" to "shared construct" reflects an important change in meaning across levels. Ultimately, questions about whether emergent effects reflect increased group-mean reliability, or whether they reflect fundamental changes in meaning (or some combination of the two) must be guided and informed by theory. As this point, empirical tools do not differentiate changes in reliability from other types of change. As we note later, understanding the nature of changes in meaning across levels is a key challenge in multilevel analyses.

In our simulation, the theoretical construct (rooms that differ in terms of temperature) is a group-level construct. Measures of this construct, however, are provided by individuals who turn out to be relatively unreliable thermometers (15% accurate). By aggregating a large number of unreliable individual measures, we change the meaning of the group measure. To elaborate, if a single individual states that a room is 82°, but 85% of the response is error, it is difficult to know whether the room was really 82° and/or whether another rater would also provide a value near 82. With one rating, it would be logical to conclude that the measure is largely error. If, on the other hand, 10 group members provide ratings and each rating is 15% accurate the resulting estimate will be more reliable. With a group mean from 10 individuals, it would be logical to assume another group of 10 would provide a similar response, and hence the estimate is more likely to reflect the true temperature. The logic underlying our example is analogous to the theory of composites which underlies scale construction from survey items in individual-level psychometrics. Indeed, as we discuss later a measure of group-mean reliability, the ICC(2), can be estimated by applying the Spearman-Brown formula to the ICC(1) and group size (see Bliese 2000).

We conclude these ideas by noting that theory surrounding emergent effects is under-developed. There continue to be opportunities to consider how emergence effects across levels reflect changes in meaning, and how group-based psychometrics can be modified and developed to help establish the construct validity of aggregate constructs. Sampson (2003) made a similar point with respect to sociological measures of collective efficacy stating that "[t]he basic idea is to take the measurement of ecological properties and social processes as seriously as we have always taken individual-level differences" (p. S57; see also Sampson et al. 1997). While several authors have expanded upon the need to more fully develop measurement in aggregate variables (e.g., Bliese et al. 2007; Chen et al. 2004), we believe construct validation of group-level variables remains as one of the most important gaps in multilevel research.

Group-Mean Reliability

As a matter of best practice, researchers often report the group-mean reliability, also known as the ICC(2), when discussing the multilevel properties of constructs—particularly constructs used as predictors. The ICC(2), like the ICC(1), can be estimated from an ANOVA model or as noted above it can also be estimated using the Spearman-Brown formula applied to the ICC(1) and group size (Bliese 2000). Take-away point seven focuses on how ICC(2) values impact researchers' ability to detect emergent effects.

Take-away point 7: Substantial ICC (2) values are not necessary for identifying emergent group-level effects (but they help).

In the simulation data, the Y variable has an ICC(1) of 0.15 and an ICC(2) of 0.64 while the ICC(1) and ICC(2) values for our X variable are 0.16 and 0.65, respectively. By conventional criteria, we would conclude that the average measures of temperature obtained by the 10 raters are not highly reliable (based on the somewhat controversial 0.70 criterion). The ICC(2) values indicate that the observed mean score values may not be entirely reproducible if re-measured.

Intuitively, it makes sense we would need reliable mean differences to detect emergent effects (e.g., Bliese 1998). Indeed, we can easily show that the strength of the emergent effect is a function of the reliability of the means. In the simulation, if we make group sizes 100 instead of 10 (using the same

random seed as before), the ICC(2) values for Y and X increase to 0.95, the group-level correlation increases to 0.94, and the *t* value associated with the mixed-effects model for X.G with both X and X.G in the model increases to 26.9.

What may not be as intuitive, however, is that we can still detect emergent effects even when ICC(2) values are low by common standards. For instance, if we run the simulation with group sizes of 2 (using the same random seed as before), we now have ICC(2) values of .34 for both X and Y, which would be considered small by conventional standards. With dyads, the observed group-mean correlation is .24 and the mixed-effects model test of emergent effects returns a *t* value of 2.07 with a parameter estimate of .29 for X.G and an estimate of −.05 for X (a total group effect of .24 for X.G).

Obviously, we are demonstrating emergent effects with weak ICC(2) values in a case where the underlying correlation between group means for X and Y is very strong (1.0). Nonetheless, the example with dyads and the example with groups of 10 demonstrate that emergent effects can be detected even if ICC(2) values do not meet conventional reliability norms of .70. This point is important for researchers who work with teams and other small groups because these researchers may have excellent ICC(1) values, but due to small group sizes will almost always have ICC(2) values less than .70. In the end, low ICC (2) values can attenuate statistical power to detect effects, but statistical power is also determined by effect size, values of ICC (1), and the number of groups (Scherbaum and Ferreter 2009); therefore, editors and reviewers should consider more than ICC(2) values when determining the merits of multilevel research particularly if ICC(1) values are significant and large.

More specifically, we encourage authors, editors, and reviewers to interpret ICC(1) and the corresponding ICC(2) values within construct-specific norms. For instance, as noted ICC(1) values for strain-related measures and mental health are often less than .05 (Bliese 2006; Murray and Short 1995). In contrast, ICC(1) values associated with perceptions of leadership are often in the range of 0.15 or larger (Bliese 2006), and other constructs such as school achievement scores typically show values between 0.15 and 0.25 (Bloom et al. 1999; Hedges and Hedberg 2007). Our point is simply that the magnitude of the ICC(2) values need to be interpreted in the context of ICC(1) norms and group-size constraints.

As an aside, researchers may also want to consider ICC(1) norms if making decisions about group sizes in cases where different options are available for group sizes. For instance, it can be difficult and costly to collect customer satisfaction data. To establish reliable group means across numerous sites, a researcher may first want to determine typical ICC(1) values on customer satisfaction scores and use that information to determine how many customers per site would need to be surveyed to produce acceptable ICC(2) values.

Organizational Data

Our third section examines five of our seven take-away points (and adds two more) using a large, organizational dataset analyzed in Bliese and Halverson (1996). The dataset (referred to as bh1996 in R) contains responses from 7382 soldiers nested in 99 groups (army companies). The dependent variable is a measure of well-being (WBEING) and the three independent variables are work hours (HRS), leadership consideration (LEAD), and cohesion (COHES). In the dataset, all variables have a group-mean centered variant (W.BEING, W.HRS, W.LEAD, W.COHES) and each have a group-mean variant (G.WBEING, G.HRS, G.LEAD, G.COHES). Code in the appendix can be used to follow the analyses.

The dependent variable of WBEING has an ICC (1) value of 0.043 indicating that 4.3% of the variance in an individual soldier's well-being measure can be explained by group membership. While the ICC (1) value is relatively small, a likelihood ratio test comparing models with and without random intercepts indicates that the ICC (1) is statistically significant (likelihood ratio of 188.83 on 1 DF, $p < .001$).

We cannot demonstrate take-away point 1 (no real downside to running mixed-effects models) because the ICC (1) is statistically significant, but we can illustrate many other points. Point 2 (biased standard errors are the result of not accounting for ICC(1) values) is evident in the top two panels of Table 7 which provide both OLS and mixed-effects model results for raw level-1 variables. Notice that OLS *t* values are too large for level-2 variables such as G.HRS (risking type I error) while OLS *t* values for level-1 variables such as HRS are too small (risking type II error), compared to mixed-effects estimates. Parameter estimates for OLS and mixed-effects models also differ because these data have unequal group sizes, but the patterns involving standard errors in the simulated data are evident in the organizational data.

We can illustrate point 3 (the interpretation of level-2 analogues of level-1 variables depends on centering choices) contrasting the upper and lower right panels of Table 7. Notice that in the lower right panel of Table 7 the parameter estimate for G.HRS is −.141 (*t* value = −7.706) representing the total relationship between average group work hours and average group well-being. In both the upper and the lower right panels of Table 7 the level-1 effect is −.026. The upper right panel shows that the difference between −.026 and −.141 (−.115) is statistically significant and provides evidence of an emergent group effect.

Point 4 (group-mean centering changes meaning to reflect relative position) is conceptual, but the lower panels in Table 7 would provide a test of a hypothesis that "an individual's work hours relative to his or her group mean are negatively related to well-being relative to his or her group." Point 5 (group-level variables predict group means of the outcome) can be illustrated by estimating an OLS model based on the

Table 7 Results of predicting soldier well-being in OLS and mixed-effects random intercept models in bh1996 data

	OLS model			Mixed-effects model			
	Estimate	Std. error	*t* value	Estimate	Std. error	*DF*	*t* value
(Intercept)	3.676	0.197	18.624	3.530	0.304	7280	11.601
HRS	−0.026	0.005	−5.677	−0.026	0.004	7280	−5.717
LEAD	0.471	0.014	32.750	0.471	0.014	7280	32.983
COHES	0.080	0.012	6.564	0.080	0.012	7280	6.610
G.HRS	−0.128	0.013	−10.135	−0.115	0.019	95	−6.127
G.LEAD	−0.234	0.049	−4.783	−0.224	0.069	95	−3.262
G.COHES	−0.031	0.063	−0.489	−0.038	0.089	95	−0.425
(Intercept)	3.676	0.197	18.624	3.530	0.304	7280	11.601
W.HRS	−0.026	0.005	−5.677	−0.026	0.004	7280	−5.717
W.LEAD	0.471	0.014	32.750	0.471	0.014	7280	32.983
W.COHES	0.080	0.012	6.564	0.080	0.012	7280	6.610
G.HRS	−0.153	0.012	−13.028	−0.141	0.018	95	−7.706
G.LEAD	0.237	0.047	5.073	0.247	0.067	95	3.677
G.COHES	0.049	0.062	0.799	0.042	0.088	95	0.479

Note: $N = 7382$ individuals in 99 groups. Upper panels based on raw data. Lower panels based on group-mean centered data

99 group means and regressing average well-being (G.WBEING) on group means for work hours, leadership, and cohesion (G.HRS, G.LEAD, G.COHES). The code to estimate this model is presented in the appendix. The estimates from the OLS model and 99 group means (not shown in the table) return coefficients of −0.130, 0.258, and 0.033 for G.HRS, G.LEAD, and G.COHES. The values from Table 7 (lower right panel) are −0.141, 0.247, and 0.042. Unlike the simulation, the OLS model and the mixed-effects model do not provide identical estimates because group sizes are unequal (the OLS model of 99 group means weights each group equally even though group sizes differ).

We cannot formally test point 6 (that emergent effects often involve changes in the meaning of variables), but we can discuss the idea from the perspective of the emergent relationships evident in the organizational data. The top right panel of Table 7 shows that emergent effects are present for leadership in addition to work hours and we discuss the effects related to leadership.

The specific form of the effect for leadership indicates significantly weaker effects at the group-level. Notice in the top right panel that the individual coefficient for LEAD is 0.471 while the "emergent" effect associated with G.LEAD is −0.224 indicating the slope between average ratings of leadership and average well-being is significantly weaker than the slope at level-1. Importantly, the coefficient of −0.224 for G.LEAD is interpreted in comparison to the slope of LEAD at the individual level, so the total slope between G.LEAD and average well-being is positive and significant at 0.247 [0.472 + (−0.224) see also the bottom right panel of Table 7]. In the end, the results suggest that individual ratings of leadership are positively correlated with individual well-being, but the positive relationship is less pronounced when examining average leadership ratings and average well-being.

Our discussion around point 6 implies that emergent group-level properties carry different meaning from their level-1 analogs. Changes in meaning can only be inferred from the emergent effects, but our example illustrates the potential. Both leadership and well-being have two analogues. Presumably, aggregate leadership with an ICC(2) of 0.928 reflects a highly reliable shared climate created by the actions and behaviors of group leaders, and aggregate well-being with an ICC(2) of 0.772 reflects some portion of health-related symptoms that respond to contagion effects from peers or shared environmental conditions. On a conceptual basis, one of the analyses is examining the effects of leadership climate on the shared contagion aspect of well-being. The other analysis is examining the effects of an individual's perceptions of leadership presumably developed through individual interactions with leaders with the individual's well-being. Results suggest that different effect sizes are associated with these two complementary views potentially suggesting that individual-based, dyadic models of leadership may be particularly important with respect to well-being.

Finally, we can illustrate the effects of decreasing the ICC(2) discussed as point 7 using Table 8. The upper

part represents an analysis of the full data set (N_{ind} = 7382; N_{groups} = 99). Notice that the raw correlation between work hours and well-being is −0.163 while the group-level correlation is −0.712 and the within (group-mean centered) is −0.111 (clear evidence of an emergent effect). The lower panel reflects results from a smaller data set created by randomly selecting 2000 observations from the full set. Notice that the ICC(2) values are attenuated (e.g., from 0.917 to 0.753 for work hours) because of smaller group sizes. The ICC(1) and raw correlations are relatively unchanged, but the between-group correlation in the small dataset between average work hours and average well-being is now −0.545. Parenthetically, part of our ability to attenuate the correlation so dramatically is because the within-correlation is very small relative to the group-level correlation. Our data behave much like the simulation of pure group-level effects because group-level effects were strong and the individual effects were weak.

Cross-Level Interactions

Our organizational data provides the opportunity to further illustrate two points related to cross-level interactions. The first is take-away point number eight:

Take-away point 8: It is helpful and informative to test and report whether level-1 slopes randomly vary among groups prior to conducting a test of a cross-level interaction, but the results of such tests should not prevent subsequently testing cross-level interactions.

An attractive feature of multilevel models is that group properties can be used to explain differences in level-1 slope coefficients as well as intercepts. For example, Morrison et al. (2011) hypothesized that the relationship between individual identification with a group and their willingness to exercise voice was stronger in groups with a favorable voice climate. Similarly, Whitener (2001) examined whether the relationship between individuals' perceived organizational support and trust in management was stronger in units with commitment-based (rather than control-based) human resources practices.

Using the bh1996 data, we can make the assumption that we expect the relationship between individual reports of work hours (HRS) and individual reports of well-being (WBEING) to be moderated by average levels of cohesion in the group (G.COHES). More specifically, we might propose that working long hours in a highly cohesive group is associated with an increase in well-being, but working long hours in a group with low cohesion is associated with a decrease in well-being. Figure 1 shows the individual-level regression lines for the first 25 groups in the bh1996 data. The slopes appear to differ. For instance, group 1 has a slope of −0.11 and group 13 has a slope of 0.03. Our question of interest is whether the overall levels of cohesion in a group (modeled as a level-2 variable) explain why the level-1 slopes differ.

An important point in the process is to try and determine whether the variability among slopes represents anything more than random error. It is entirely possible that the slope variability we observe in Fig. 1 represents random variability around a true correlation of −0.16. Indeed, in Fig. 2, we simulate 25 "groups" by generating 75 pairs (average group size in the real data) of random numbers around a true correlation of −0.16. Notice that the slope variability is quite similar to the variability we observe in the real data (Fig. 1). Formally, we can test whether slopes significantly vary by contrasting −2log likelihood values for (a) a model with a random intercept with (b) a model with a random intercept and random slope. If the model with random slope provides a better fit, we have evidence to suggest that slope differences are more than random error. Some programs (e.g., HLM) also provide variance estimates of the slope parameter which can be used to test significance of the slope variance term.

A formal test of whether the slopes vary across groups is presented in the appendix. The −2log likelihood ratio is 3.12 with a p value of .21 (*ns*). The likelihood test is known to be conservative. LaHuis and Ferguson (2009) recommend treating the test as a one-tailed test, but even a p value of

Table 8 Raw, within-group, and between-group correlations in the full bh1996 data and a random subset

	ICC(1)	ICC(2)	Raw *r*	*r* (between group)	*r* (within group)
Full data set					
Hours	.129	.917	−.163	−.712	−.111
Well-being	.043	.772			
Random subset of individuals from each group					
Hours	.131	.753	−.161	−.545	−.106
Well-being	.047	.499			

Note: N = 7382 individuals in 99 groups (full data set). N = 2000 individuals in 99 groups (ramdom subset)

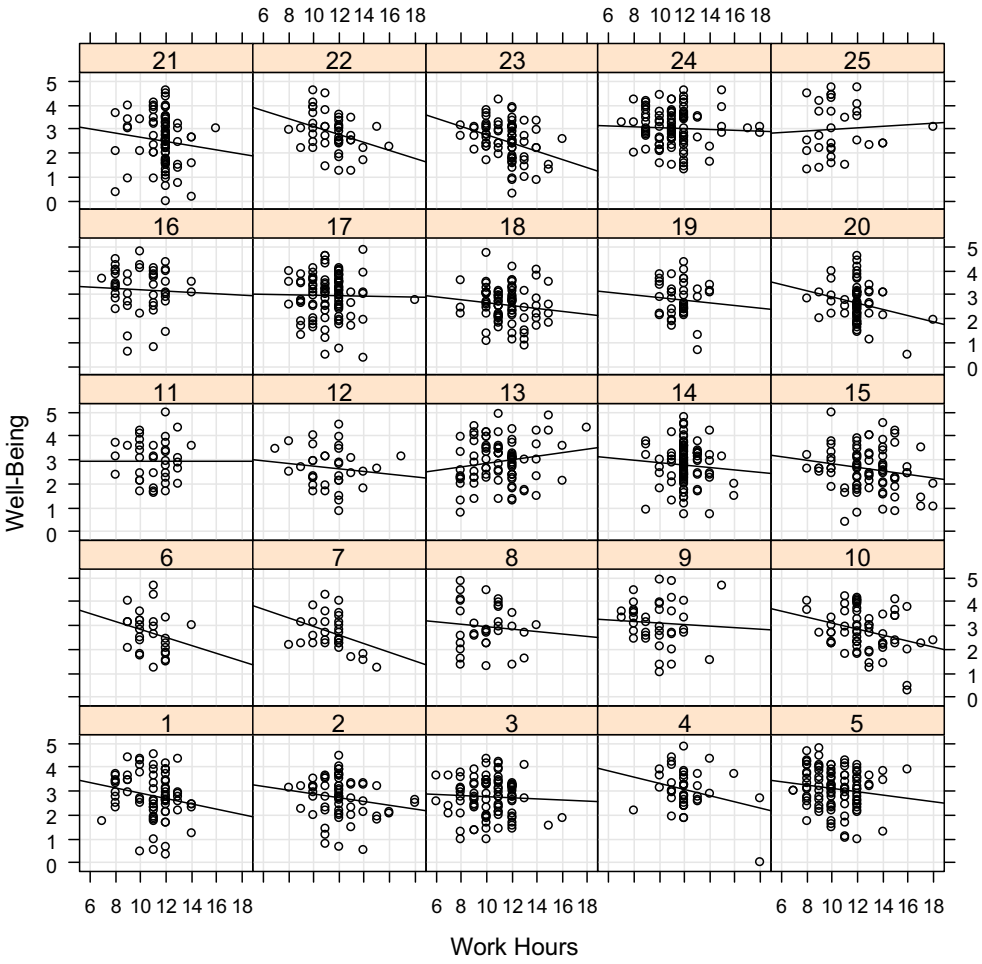

Fig. 1 Relationship between ratings of work hours and well-being for the first 25 groups in the bh1996 dataframe

.105 would provide little evidence of significant variability. In other words, the test would lead us to conclude that the observed slope differences in Fig. 1 likely represent normal variability rather than meaningful group differences. Despite these results, if we had a strong theoretical backing for the hypothesized interaction, we could choose to proceed with the test.

In practice, the first author has seen many cases in the review process where researchers omit reporting tests of slope variability when examining cross-level interactions. To some degree, the omission is understandable because authors such as Snijders and Bosker (1999) and others make a strong case that cross-level interactions should still be tested even if the slopes do not statistically vary if theory supports conducting such tests. Ultimately, though, we believe that routinely providing slope variability information helps readers understand the nature of the data and we encourage editors and reviewers to routinely request this information.

In our final take-away point, we refer to our earlier discussion of group-mean centering where we indicated the importance of using group-mean centering to verify results involving cross-level interactions. Therefore, take-away point 9 is:

Take-away point 9: When testing cross-level interactions with raw or grand-mean centered variables, it is important to use group-mean centering to verify model results.

We use the organizational data to elaborate on this point. The upper panel in Table 9 shows a model based on raw data, and the t value of −2.211 suggests a statistically significant cross-level interaction. The problem with interpreting the t value of −2.211 is that this apparent cross-level interaction could be a function of group means interacting. As noted by Raudenbush and Bryk (2002), Hofmann and Gavin (1998), and others, group-level moderators occasionally interact with the group-mean of the level-1 predictor rather than the level-1 slope between predictor and outcome. When group-mean interactions exist, models based on raw or grand-mean centered variables can erroneously indicate a cross-level interaction.[6]

Given the potential confound with group-means, it is helpful to verify cross-level interactions based on raw or grand-mean centered data with a test using a group-mean centered

[6] Interested readers can execute the final simulation in the appendix to see how a group-mean interaction can be detected in a cross-level interaction model and then how group-mean centering corrects the problem.

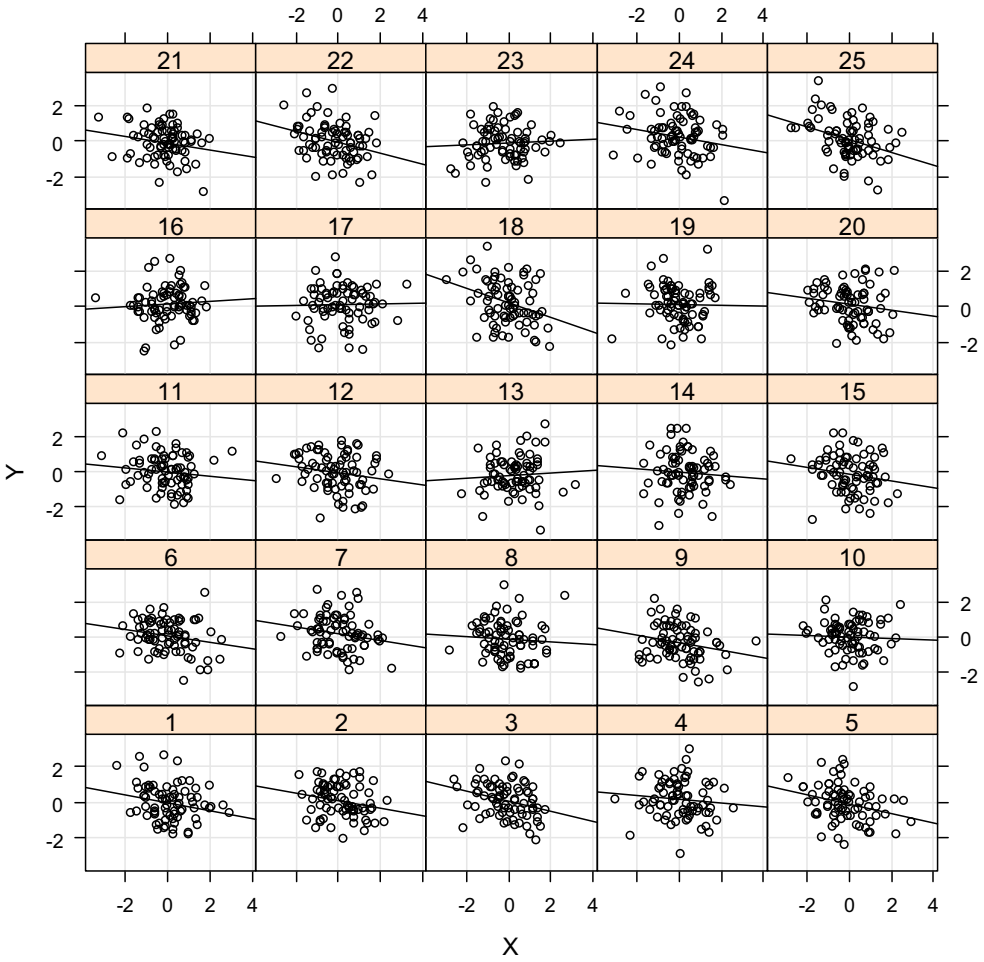

Fig. 2 Simulated data for 25 groups of 75 with rho of −.16

level-1 variable. The lower part of Table 9 suggests that a group-mean interaction effect is inflating our apparent cross-level interaction because the effect is no longer significant when using group-mean centered variables (t value of −1.567). Practically speaking, results of this nature would be problematic because our theory would almost certainly have been based on the premise that the strength of the individual-level relationship between work hours and well-being was being moderated by group cohesion and authors should then acknowledge that their hypotheses did not receive support. Group-mean centered variables sometimes weaken effects (as they did here), but we have also seen effects strengthened and in many cases models differ only in minor ways. Even when results do not differ, we urge authors to conduct this test, and add report the outcome in a footnote.

Table 9 Results of a cross-level interaction with a raw work hours (upper panel) and group-mean centered work hours (lower panel) in bh1996 data

	Value	Std. error	*DF*	*t* value	*p* value
(Intercept)	0.750	0.821	7281	0.914	0.361
HRS	0.098	0.068	7281	1.437	0.151
G.COHES	0.858	0.268	97	3.196	0.002
HRS:G.COHES	−0.049	0.022	7281	−2.211	0.027
(Intercept)	1.812	0.298	7281	6.075	0.000
W.HRS	0.063	0.070	7281	0.900	0.368
G.COHES	0.314	0.097	97	3.234	0.002
W.HRS:G.COHES	−0.036	0.023	7281	−1.567	0.117

Note: N = 7382 individuals in 99 groups

Discussion

In summary, the nine points we highlight offer specific guidance to researchers regarding several areas of multi-level research. Our simulations, examples, and take-away points are designed to reduce ambiguity about certain aspects of mixed-effects models and help researchers use the methods most effectively. We can roughly arrange the points around the themes of (a) details surrounding model use, (b) how centering changes models, and (c) how results should be interpreted. Table 10 provides a summary and restatement of the points. Below we discuss the points in more detail.

Table 10 Summary of points

Details about when and how to use mixed-effects models
Point 1: Mixed-effects models can be applied to nested data, even if no group-level effects are expected or evident
Point 2: Failing to use mixed effects to model nested data can result in too liberal Type I Errors (for group-level effects) and too conservative Type II Errors (for lower-level effects)
Point 7: Substantial ICC(2) values are not a prerequisite for identifying group-level effects (but they help)
Point 8: In models involving cross-level interactions, best practice is to report information about the variability of slopes before estimating models that predict slopes
How does centering variables change models?
Point 3: Centering choices of level-1 variables change the meaning of level-2 analogues of the same variable. Use raw (or grand-mean centered) if testing whether level-2 slopes differ from level-1 slopes. Use group-mean centering if testing whether the level-2 relationship is related to the group average of the dependent variable
Point 4: Group mean centering changes the meaning of the level-1 construct to where the term reflects relative position in a group
Point 9: Group mean centering level-1 variables is important for verifying cross-level interactions
How do I interpret my mixed-effects model results?
Point 5: Variables at level-2 predict group means of lower-level outcomes
Point 6: Emergent effects where slopes at level-2 are significantly different from slopes at level-1 suggest some fundamental change in meaning of variables across levels

Details About Model Use Several of our main points clarify misconceptions that might have caused researchers to be reticent to utilize mixed-effects models. Point 1 recommends that researcher default to using mixed-effects models whenever they have nested data because there is virtually no penalty for using mixed-effects models where OLS would have sufficed, whereas misapplying OLS can lead to inaccurate results. Point 2 elaborated on this recommendation by showing that using OLS to model non-independent data leads to type I errors (for group-level effects) and type II error (for lower-level effects). Point 7 demonstrated that substantial ICC(2) values are not a prerequisite for identifying group-level effects, suggesting that low ICC(2) values for level-2 predictors should not deter researchers from using mixed-effects models. Finally, point 8 emphasizes that tests for significant variance in slopes should routinely be reported.

How Centering Changes Models The points in this article clarify differences between group mean centering and grand-mean centering. Point 3 showed how the parameter estimate of a level-2 analogue of a level-1 variable changed meaning depending upon the centering of the level-1 variable such that (a) formal tests of emergent effects at level-2 are most easily accomplished using raw (or grand-mean) variables and (b) formal tests of whether the level-2 variable is related to the group average of the dependent variable is most easily accomplished by group-mean centering the level-1 variable. Point 4 stated that if group-mean centering is used, hypotheses should be explicit about referring to individuals' relative position within their group. Finally, point 9 emphasized that group-mean centering is an important supplemental test when examining cross-level interactions.

Interpreting Results Many points in this article revolve around interpretation; however, we highlight two additional points. First, in point 5, we clarify that level-2 independent variables predict group means of the outcome variable. Second, in point 6, we discuss that when relationships between group means differ significantly from relationships between individual variables (i.e., an emergent effect is present), the results suggest that the group-level constructs have undergone fundamental change in meaning.

Conclusion

As theories in organizational behavior and organizational psychology address new multilevel issues, and multilevel data become increasingly available to researchers, we anticipate that mixed-effects modeling will remain widely used. Based on our experiences with using and evaluating mixed-effects models, we have presented nine key points we believe could enhance the utility of the models. Little here is new to methodology experts, but we hope the take-away points will serve as a convenient reference for those wishing to gain more familiarity with the approach. We also hope that some of the principles we have laid out will provide a basis for challenging some of the existing norms in the analysis of multilevel data.

Appendix: R Code

```
#######################################################
# Set up Independent data and assign Group IDs to observations #
#######################################################
set.seed(12532)
Y<-rnorm(1000)
X<-.30*Y+sqrt(1-.30^2)*rnorm(1000)
cor(X,Y)
G.ID<-rep(1:100,10)
NO.GRP<-data.frame(G.ID=G.ID,Y=Y,X=X)

# Calculate group means for X. Assign means back to members as X.G
TDAT<-aggregate(NO.GRP[,c("G.ID","X")],list(NO.GRP$G.ID),mean)
NO.GRP<-merge(NO.GRP,TDAT[,c("G.ID","X")],by="G.ID",
suffixes=c("",".G"))

# Table 1
NO.GRP[1:15,]

# Load library for ICC functions and link to nlme functions. Note:
# It may be necessary to install the multilevel package from CRAN
# using the "Install package(s)" option under "Packages" in R prior
# to issuing the library command below.
library(multilevel)

# ICC(1) estimate via mixed-effects
tmod<-lme(fixed=Y~1,random=~1|G.ID,data=NO.GRP)
VarCorr(tmod)

#> 4.299649e-09/(4.299649e-09+9.986574e-01)
#[1] 4.305429e-09

# ICC(1) estimate via ANOVA
tmod<-aov(Y~as.factor(G.ID),data=NO.GRP)
ICC1(tmod)

#Table 2 (upper panel)
tmod<-lm(Y~X+X.G, data=NO.GRP)
summary(tmod)$coef

#Table 2 (lower panel)
tmod<-lme(fixed=Y~X+X.G,random=~1|G.ID,data=NO.GRP)
summary(tmod)$tTable

#################
# Re-run simulation with data that returns a slightly positive ICC(1)
#################
set.seed(125321)
Y<-rnorm(1000)
X<-.30*Y+sqrt(1-.30^2)*rnorm(1000)
cor(X,Y)
```

```
G.ID<-rep(1:100,10)
NO.GRP<-data.frame(G.ID=G.ID,Y=Y,X=X)

# ICC(1) estimate via mixed-effects
tmod<-lme(fixed=Y~1,random=~1|G.ID,data=NO.GRP)
VarCorr(tmod)

# ICC(1) estimate via ANOVA
tmod<-aov(Y~as.factor(G.ID),data=NO.GRP)
summary(tmod)
ICC1(tmod)

# Calculate group means for X. Assign means back to members as X.G
TDAT<-aggregate(NO.GRP[,c("G.ID","X")],list(NO.GRP$G.ID),mean)
NO.GRP<-merge(NO.GRP,TDAT[,c("G.ID","X")],by="G.ID",
suffixes=c("",".G"))

#Table 3 (upper panel)
tmod<-lm(Y~X+X.G, data=NO.GRP)
summary(tmod)$coef

#Table 3 (lower panel)
tmod<-lme(fixed=Y~X+X.G, random=~1|G.ID,data=NO.GRP)
summary(tmod)$tTable

# Create a group-mean centered variable (W.X)
NO.GRP$W.X<-NO.GRP$X-NO.GRP$X.G

#Table 4 (upper panel)
tmod<-lm(Y~W.X+X.G, data=NO.GRP)
summary(tmod)$coef

#Table 4 (lower panel)
tmod<-lme(fixed=Y~W.X+X.G,random=~1|G.ID,data=NO.GRP)
summary(tmod)$tTable

#################
# Create a group-mean dataset using the simulated data with
# a slight positive ICC(1)
#################
TDAT<-aggregate(NO.GRP[,c("Y","X")],list(NO.GRP$G.ID),mean)
names(TDAT)<-c("G.ID","Y.G","X.G")

#Table 5 (results from 100 group means)
tmod<-lm(Y.G~X.G, data=TDAT)
summary(tmod)$coef

#Add another group-level predictor, Z.G, with no level-1 analogue
set.seed(125321)
Y<-rnorm(1000)
X<-.30*Y+sqrt(1-.30^2)*rnorm(1000)
G.ID<-rep(1:100,10)
NO.GRP<-data.frame(G.ID=G.ID,Y=Y,X=X)
```

```
NO.GRP<-NO.GRP[order(NO.GRP$G.ID),]

NO.GRP$Z.G<-rep(rnorm(100),each=10)

tmod<-lme(fixed=Y~Z.G, random=~1|G.ID, data=NO.GRP)
summary(tmod)$tTable

TDAT<-aggregate(NO.GRP[,c("Y","Z.G")],list(NO.GRP$G.ID),mean)
names(TDAT)<-c("G.ID","Y.G","Z.G")
tmod<-lm(Y.G~Z.G, data=TDAT)
summary(tmod)$coef

#################
# Simulate pure group-level data: This simulation relies on the
# sim.icc function in the multilevel library.
#################
library(multilevel)

set.seed(632342)
ALL.GRP<-sim.icc(gsize=10,ngrp=100,icc1=.15,nitems=2)
names(ALL.GRP)<-c("G.ID","Y","X")

# covariance theorem decomposition results
with(ALL.GRP, waba(x=X, y=Y, grpid=G.ID))

# Randomly sort X on a group-by-group basis into X2 (footnote 5)
set.seed(467432)
ALL.GRP$X2<-unlist(tapply(ALL.GRP$X,ALL.GRP$G.ID,sample))
ALL.GRP[1:10,] #observe that X2 values are X values in new order
with(ALL.GRP,waba(Y,X2,G.ID))

#Estimate the ICC(1) and ICC(2) values for the Y variable using
#the ANOVA method
tmod<-aov(Y~as.factor(G.ID),data=ALL.GRP)
ICC1(tmod)
ICC2(tmod)

# Calculate group means for X. Assign means back to members as X.G
TDAT<-aggregate(ALL.GRP[,c("G.ID","X")],list(ALL.GRP$G.ID),mean)
ALL.GRP<-merge(ALL.GRP,TDAT[,c("G.ID","X")],
by="G.ID",suffixes=c("",".G"))

#Table 6 (upper panel)
tmod<-lm(Y~X+X.G,data=ALL.GRP)
summary(tmod)$coef

#Table 6 (lower panel)
tmod<-lme(fixed=Y~X+X.G, random=~1|G.ID, data=ALL.GRP)
summary(tmod)$tTable

#Create a group-mean centered variable (W.X)
ALL.GRP$W.X<-ALL.GRP$X-ALL.GRP$X.G
#Estimate total effect of X.G with a group-mean centered X variable
```

```
tmod<-lme(fixed=Y~W.X+X.G, random=~1|G.ID, data=ALL.GRP)
summary(tmod)$tTable

# Estimate ICC(1) and ICC(2) values for both X and Y using
# ANOVA method
tmod<-aov(Y~as.factor(G.ID),data=ALL.GRP)
ICC1(tmod)
ICC2(tmod)

tmod<-aov(X~as.factor(G.ID),data=ALL.GRP)
ICC1(tmod)
ICC2(tmod)


#################
# Demonstrate the effect of alternative group sizes on ICC(2)
# values, and on the ability to detect emergent effects.
# Substitute either 2 or 100 for gsize in sim.icc function.
#################

set.seed(632342)

ALL.GRP<-sim.icc(gsize=2,ngrp=100,icc1=.15,nitems=2)
names(ALL.GRP)<-c("G.ID","Y","X")
ALL.GRP[1:15,]
with(ALL.GRP,waba(x=X, y=Y, grpid=G.ID))

tmod<-aov(Y~as.factor(G.ID),data=ALL.GRP)
ICC1(tmod)
ICC2(tmod)

tmod<-aov(X~as.factor(G.ID),data=ALL.GRP)
ICC1(tmod)
ICC2(tmod)

TDAT<-aggregate(ALL.GRP[,c("G.ID","X")],list(ALL.GRP$G.ID),mean)
ALL.GRP<-merge(ALL.GRP,TDAT[,c("G.ID","X")],
by="G.ID",suffixes=c("",".G"))

tmod<-lme(fixed=Y~X+X.G, random=~1|G.ID, data=ALL.GRP)
summary(tmod)$tTable

#################
# Load organizational data: The bh1996 (Bliese & Halverson 1996)
# dataset from the multilevel library
#################
library(multilevel)
data(bh1996) #bring data from library to workspace

#Estimate ICC(1) using mixed-effects model
null<-lme(fixed=WBEING~1,random=~1|GRP,data=bh1996)
VarCorr(null)
0.03580077/(0.03580077+0.78949727)
```

```
#[1] 0.0433792

#likelihood test of significance of ICC(1)
null.gls<-gls(WBEING~1,data=bh1996)
anova(null.gls,null)

#Table 7 (upper left panel): OLS raw variables.
tmod<-lm(WBEING~HRS+LEAD+COHES+G.HRS+G.LEAD+G.COHES, data=bh1996)
summary(tmod)$coef

#Table 7 (upper right panel): Mixed-effect raw variables.
tmod<-lme(fixed=WBEING~HRS+LEAD+COHES+G.HRS+G.LEAD+G.COHES, random=~1|GRP, data=bh1996)
summary(tmod)$tTable

#Table 7 (lower left panel): OLS group-mean centered
tmod<-lm(WBEING~W.HRS+W.LEAD+W.COHES+G.HRS+G.LEAD+G.COHES,
data=bh1996)
summary(tmod)$coef

#Table 7 (lower right panel): Mixed-effect group-mean centered
tmod<-lme(fixed=WBEING~W.HRS+W.LEAD+W.COHES+G.HRS+G.LEAD+G.COHES,
random=~1|GRP, data= bh1996)
summary(tmod)$tTable

# Estimate ICC values for leadership and well-being
mult.icc(bh1996[,c("LEAD","WBEING")],bh1996$GRP)

# Estimate a group mean analysis
TDAT<-aggregate(bh1996[,c("WBEING","HRS","LEAD","COHES")],
list(bh1996$GRP),mean,na.rm=TRUE)
summary(lm(WBEING~HRS+LEAD+COHES,data=TDAT))$coef

#################
# Change ICC(2) values
#################

# Table 8 values
mult.icc(bh1996[,c("HRS","WBEING")],bh1996$GRP)
round(with(bh1996,waba(x=WBEING,y=HRS,grpid=GRP))$Cov.Theorem,dig=4)

set.seed(271843)
bh1996.small<-bh1996[sample(1:7382, 2000),]
length(unique(bh1996.small$GRP))
mult.icc(bh1996.small[,c("HRS","WBEING")],
bh1996.small$GRP)
round(with(bh1996.small,waba(x=WBEING,y=HRS,grpid=GRP))$Cov.Theorem,
dig=4)

#################
# Figure 1, Figure 2 and Table 9
#################
library(lattice)
# Code for Figure 1
```

```
xyplot(WBEING~HRS|as.factor(GRP),data=bh1996[1:1582,],
subset=HRS>5&HRS<19,
type=c("p","g","r"),col="black",col.line="black",
xlab="Work Hours",
ylab="Well-Being")

# Find group-specific OLS based intercepts and slopes
lmList(WBEING~HRS|GRP,bh1996)

# Function to generate correlations on a group by group basis.
# The function is used below to generate data for Figure 2.
tfun<-function(gsize,ngroup,rho){
OUTPUT<-data.frame(GRP=rep(NA,gsize*ngroup),X=rep(NA,gsize*ngroup),
Y=rep(NA,gsize*ngroup))
for(i in 1:ngroup){
Y<-rnorm(gsize)
X<-rho*Y+sqrt(1-rho^2)*rnorm(gsize)
OUTPUT[((i * gsize - gsize) + 1):(i *gsize),]<-data.frame(GRP=i,X=X,Y=Y)
}
return(OUTPUT)
}

# Figure 2: Simulate the bh1996 Work Hour and Well-Being
# relationship for 25 groups each with 75 members.
set.seed(12532)
SIM.DAT<-tfun(75,25,-.16)

xyplot(Y~X|as.factor(GRP),SIM.DAT,
type=c("p","g","r"),col="black",col.line="black")

#################
# Test whether slopes randomly vary in bh1996 data
#################
tmod<-lme(fixed=WBEING~HRS, random=~1|GRP, data=bh1996)
tmod2<-lme(fixed=WBEING~HRS, random=~HRS|GRP,data=bh1996)
anova(tmod,tmod2)

#> anova(tmod,tmod2)
# Model df AIC BIC logLik Test L.Ratio p-value
#tmod 1 4 19249.79 19277.42 -9620.896
#tmod2 2 6 19250.67 19292.11 -9619.337 1 vs 2 3.118139 0.2103

# Table 9(upper panel): Cross-level interactions and group-mean
# Centering variables
tmod<-lme(fixed=WBEING~HRS*G.COHES, random=~1|GRP, data=bh1996)
summary(tmod)$tTable

# Table 9 (Lower panel):
tmod<-lme(fixed=WBEING~W.HRS*G.COHES, random=~1|GRP, data=bh1996)
summary(tmod)$tTable

#################
# Set up a simulation to demonstrate how an interaction
```

```
# involving group means can erroneously be identified as a
# cross-level interaction and how group-mean centering then
# corrects the problem. Conceptually, while the code has numerous
# steps, it functionally just creates 100 group means where 50
# group means have a positive X,Y relationship and 50 group means
# have a negative X,Y relationship. The 100 group means are
# transformed into individual variables by replicating each group
# mean 10 times and adding random error. Thus there are no
# individual-level relationships except those associated with the
# group mean. Z.G is a group-level variable where a value of 0 is
# associated with group-means having a positive relationship, and
# a value of 1 is associated with the group means having a negative
# relationship.
#################
set.seed(2500852)

# Create 50 X and Y values to use as group means with
# a POSITIVE correlation (rho) of .50
Y.1<-rnorm(50)
X.1<-.50*Y.1+sqrt(1-.50^2)*rnorm(50)
cor(X.1,Y.1)

# Create 50 X and Y values to use as group means with
# a NEGATIVE correlation (rho) of -.50
Y.2<-rnorm(50)
X.2<-(-.50)*Y.2+sqrt(1-(-.50)^2)*rnorm(50)
cor(X.2,Y.2)

# Combine the 50 values into a single vector with
# 100 values (half of which are positive and half of
# which are negative)
Y<-c(Y.1,Y.2)
X<-c(X.1,X.2)

# Replicate the 100 values 10 times each for a vector of
# length 1000
Y<-rep(Y,each=10)
X<-rep(X,each=10)

# Add some random error to each of the 1000 values
Y<-Y+rnorm(1000,sd=.1)
X<-X+rnorm(1000,sd=.1)

# Combine the X and Y vectors into a data.frame add Z.G
# as a group-level variable. When Z.G=0 the X and Y group mean
# values were positively correlated; when Z.G=1 X and Y values were
# negatively correlated
GMEAN.INT<-data.frame(Y=Y,X=X,Z.G=rep(0:1,each=500),
GRP=rep(1:100,each=10))

# Estimate a cross-level interaction model omitting the random
# slope for X because no slope variability exists and
# include X in random=~X|GRP causes convergence problems.
```

```
tmod<-lme(Y~X*Z.G,random=~1|GRP,data=GMEAN.INT)
summary(tmod)$tTable

# Create a group-mean variable for X to create a group-mean
# centered variable W.X
TDAT<-aggregate(GMEAN.INT$X,list(GMEAN.INT$GRP),mean)
names(TDAT)<-c("GRP","X.G")
GMEAN.INT<-merge(GMEAN.INT,TDAT,by="GRP")
GMEAN.INT$W.X<-GMEAN.INT$X-GMEAN.INT$X.G

# Estimate a cross-level interaction model using
# Group-mean centered X (W.X). Notice the cross-level interaction
# is no longer significant.
tmod<-lme(Y~W.X*Z.G,random=~1|GRP,data=GMEAN.INT)
summary(tmod)$tTable

# Show how a fixed-effect OLS model can also recover the cross
# level interaction. Z.G cannot be included as a main-effect because
# the variance is captured by the dummy codes for group, but
# the interaction term X:Z.G can be included.
tmod<-lm(Y~X+X:Z.G+as.factor(GRP),data=GMEAN.INT)
summary(tmod)$coef[c(1,2,102),]
```

References

Aguinis, H., & Molina-Azorin, J. F. (2015). Using multilevel modeling and mixed methods to make theoretical progress in microfoundations for strategy research. *Strategic Organization, 13*(4), 353–364.

Alwin, D. F. (1976). Assessing school effects: Some identities. *Sociology of Education, 49*(4), 294–303.

Alwin, D. F., & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review, 40*, 37–47.

Barrick, M. R., Stewart, G. L., Neubert, M. J., & Mount, M. K. (1998). Relating member ability and personality to work-team processes and team effectiveness. *Journal of Applied Psychology, 83*(3), 377–391.

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin, 83*(5), 762–765.

Bliese, P. D. (1998). Group size, ICC values, and group-level correlations: A simulation. *Organizational Research Methods, 1*(4), 355–373.

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

Bliese, P. D. (2006). Social climates: Drivers of soldier well-being and resilience. In A. B. Adller, C. A. Castro, & T. W. Britt (Eds.), *Military life: The psychology of serving in peace and combat (Vol. 2. Operational stress, pp. 213–234)*. Westport: Praeger Security International.

Bliese, P. D., & Halverson, R. R. (1996). Individual and nomothetic models of job stress: An examination of work hours, cohesion and well-being. *Journal of Applied Social Psychology, 26*(13), 1171–1189.

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*(4), 400–417.

Bliese, P. D., Chan, D., & Ployhart, R. E. (2007). Multilevel methods: Future directions in measurement, longitudinal analyses, and nonnormal outcomes. *Organizational Research Methods, 10*(4), 551–563.

Bloom, H. S., Bos, J. M., & Lee, S. W. (1999). Using cluster random assignment to measure program impacts: Statistical implications for the evaluation of education programs. *Evaluation Review, 23*(4), 445–469.

Chen, G., Mathieu, M. J., & Bliese, P. D. (2004). A framework for conducting multilevel construct validation. In F. Dansereau & F. J. Yammarino (Eds.), *Research in multi-level issues, volume 3: Multi-level issues in organizational behavior and processes* (pp. 273–303). Oxford: Elsevier Science.

Dansereau, F., Alutto, J. A., & Yammarino, F. J. (1984). *Theory testing in organizational behavior: The varient approach*. Englewood Cliffs Prentice-Hall.

Firebaugh, G. (1978). A rule for inferring individual-level relationships from aggregate data. *American Sociological Review, 43*(4), 557–572.

Firebaugh, G. (1980). Groups as contexts and frog ponds. In K. H. Roberts & L. Burstein (Eds.), *Issues in aggregation* (pp. 43–52). San Francisco: Jossey-Bass.

González-Morales, M. G., Peiró, J. M., Rodriguéz, I., & Bliese, P. D. (2012). Perceived collective burnout: A multilevel explanation of burnout. *Anxiety, Stress, and Coping, 25*(1), 43–61.

Hedges, L., & Hedberg, E. C. (2007). Intraclass correlation values for planning group randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60–87.

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management, 24*(5), 623–641.

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

LaHuis, D. M., & Ferguson, M. W. (2009). The accuracy of significance tests for slope variance components in multilevel random coefficient models. *Organizational Research Methods, 12*(3), 418–435.

Lang, J., Thomas, J. L., Bliese, P. D., & Adler, A. B. (2007). Job demands and job performance: The mediating effect of psychological and physical strain and the moderating effect of role clarity. *Journal of Occupational Health Psychology, 12*, 116–124.

Li, Y., Wang, Z., Yang, L.-Q., & Liu, S. (2016). The crossover of psychological distress from leaders to subordinates in teams: The role of

abusive supervision, psychological capital, and team performance. *Journal of Occupational Health Psychology, 21*(2), 142–153.

Liao, H., & Chuang, A. (2007). Transforming service employees and climate: A multilevel, multisource examination of transformational leadership in building long-term service relationships. *Journal of Applied Psychology, 92*(4), 1006–1019.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models*. Cary: SAS Institute.

Liu, D., Liao, H., & Loi, R. (2012). The dark side of leadership: A three-level investigation of the cascading effect of abusive supervision on employee creativity. *Academy of Management Journal, 55*(5), 1187–1212.

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

Luciano, M. M., Mathieu, J. E., & Ruddy, T. M. (2014). Leading multiple teams: Average and relative external leadership influences on team empowerment and effectiveness. *Journal of Applied Psychology, 99*(2), 322–331.

Mathieu, J. E., & Kohler, S. S. (1990). A cross-level examination of group absence influences on individual absence. *Journal of Applied Psychology, 75*(2), 217–220.

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods, Advance online publication*. doi:10.1037/met0000078.

Miron-Spektor, E., Erez, M., & Naveh, E. (2011). The effect of conformist and attentive-to-detail members on team innovation: Reconciling the innovation paradox. *Academy of Management Journal, 54*(4), 740–760.

Morgeson, F. P., & Hofmann, D. A. (1999). The structure and function of collective constructs: Implications for multilevel resarch and theory development. *Academy of Management Review, 24*(2), 249–265.

Morrison, E. W., Wheeler-Smith, S. L., & Kamdar, D. (2011). Speaking up in groups: A cross-level study of group voice climate and voice. *Journal of Applied Psychology, 96*(1), 183–191.

Murray, D. M., & Short, B. (1995). Intra-class correlation among measures related to alcohol use by young adults: Estimates, correlates and applications in intervention studies. *Journal of Studies on Alcohol, 56*, 681–694.

Ostroff, C. (1993). Comparing correlations based on individual-level and aggregated data. *Journal of Applied Psychology, 78*, 569–582.

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

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15*(3), 209–233.

Raudenbush, S. W. (2009). Adaptive centering with random effects: An alternative to the fixed effects model for studying time-varying treatments in school settings. *Education, 4*, 468–491.

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

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*, 351–357.

Sampson, R. J. (2003). The neighborhood context of well-being. *Perspectives in Biology and Medicine, 46*(3), S53–S64.

Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science, 277*(5328), 918–924.

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods, 12*(2), 347–367.

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

Verbeek, M. (2008). *A guide to modern econometrics*. John Wiley & Sons.

Wang, X.-H., & Howell, J. M. (2010). Exploring the dual-level effects of transformational leadership on followers. *Journal of Applied Psychology, 95*(6), 1134–1144.

Whitener, E. M. (2001). Do "high commitment" human resource practices affect employee commitment? A cross-level analysis using hierarchical linear modeling. *Journal of Management, 27*(5), 515–535.

Wolfinger, R. D. (1997). An example of using mixed models and PROC MIXED for longitudinal data. *Journal of Biopharmaceutical Statistics, 7*(4), 481–500.