Group Dynamics: Theory, Research, and Practice
2008, Vol. 12, No. 2, 127–140

1089-2699/08/$12.00 DOI: 10.1037/1089-2699.12.2.127

Assumptions of Cross-Level Measurement and Structural Invariance in the Analysis of Multilevel Data: Problems and Solutions

Michael J. Zyphur
National University of Singapore

Seth A. Kaplan
George Mason University

Michael S. Christian
University of Arizona

This article demonstrates assumptions of invariance that researchers often implicitly make when analyzing multilevel data. The first set of assumptions is measurement-based and corresponds to the fact that researchers often conduct single-level exploratory and confirmatory factor analyses, and reliability analyses, with multilevel data. The second assumption, that of structural invariance, is engineered into the common multilevel random coefficient model, in that such analyses impose structural invariance across multiple levels of analysis when lower-level relationships represent both between- and within-groups effects. The nature of these assumptions, and ways to address their tenability, are explored from a conceptual standpoint. Then an empirical example of these assumptions and ways to address them is provided.

Keywords: multilevel, invariance, measurement, cross-level, constraint

> *A frequent shortcoming when ignoring the multilevel structure of the data is not what is misestimated, but what is not learned.*—B. Muthén (1997, p. 455)

In many settings, researchers deal with data that are multilevel in nature. For example, individuals may be nested within workgroups (e.g., Hofmann, Morgeson, & Gerras, 2003) and observations over time may be hierarchically arranged within individuals (e.g., Thoresen, Bradley, Bliese, & Thoresen, 2004). This recognition has prompted the adoption of statistical techniques that allow for the efficient and unbiased analysis of these hierarchical data, such as multilevel random coefficient modeling (MRCM; i.e., hierarchical linear modeling) (see Hox, 2002; Moritz & Watson, 1998; Nezlek & Zyzniewski, 1998; Pollack, 1998; Raudenbush & Bryk, 2002). Although the use of such techniques is becoming common in many fields of study, researchers still often analyze multilevel data as if they were single-level in nature. In these cases, assumptions of parameter invariance are made across the levels of analysis that exist in the data, which we term "assumptions of cross-level invariance." Such assumptions are of two forms: cross-level measurement invariance and cross-level structural invariance.

Cross-level measurement invariance is assumed when single-level confirmatory/exploratory factor analyses (CFA, EFA) and reliability analyses are conducted with multilevel data—reliability and factor analyses describe the measurement-based properties of observed variables (Gorsuch, 1983). In these analyses, a single estimate of factor loadings, factor structure, and reliability are estimated; these estimates collapse across the multiple levels which exist in the data, yielding a single estimate for any given parameter across the multiple levels inherent in a researcher's data (for examples, see DeShon, Kozlowski, Schmidt, Milner, & Weichmann, 2004; Myers, Feltz, & Short,

Michael J. Zyphur, Department of Management and Organization, National University of Singapore Business School, National University of Singapore; Seth A. Kaplan, Department of Psychology, George Mason University; Michael S. Christian, Department of Management and Organizations, University of Arizona.

We thank Robyn Ely for her help in procuring the dataset used in the current work, without her help it would not have been possible.

Correspondence concerning this article should be addressed to Michael J. Zyphur, Department of Management and Organization, National University of Singapore Business School, 1 Business Link, National University of Singapore, Singapore 117592. E-mail: zyphurmj@yahoo.com

2004; Myers, Payment, & Short, 2004). The problems with such an approach is that (a) researchers miss the opportunity to discover similarities and differences across levels of analysis in the functioning of their observed variables, differences that could have interesting theoretical importance (see Cronbach, 1976; Sirotnik, 1980) and (b) researchers use such single-level analyses to justify the functioning of observed variables, which are then often used to investigate relationships at multiple levels of analysis.

Cross-level structural invariance is assumed when effects among variables are constrained to equality across levels of analysis—relationships among different variables are structural parameters (Vandenberg & Lance, 2000). Interestingly, this often occurs in multilevel data analyses, where researchers are explicitly attempting to account for the multilevel structure of their data (e.g., Luong & Rogelberg, 2005). Cross-level structural invariance is often unknowingly assumed because many MRCM programs have a modeling default which forces invariance on researchers' structural models across levels. This default constrains the relationships among "lower-level" variables to equality across the within- and between-groups levels of analysis or across the within- and between-individual levels of analysis (for examples, see Hofmann et al., 2003; Chen, Bliese, & Mathieu, 2005). The shortcoming of such invariance constraints is not simply that researchers may be unjustifiably imposing structural invariance across levels; more importantly, researchers miss the opportunity to discover cross-level differences in relationships (i.e., composition effects), differences that "occur with considerable regularity" (Raudenbush & Bryk, 2002, p. 140) and have implications for understanding constructs across levels (see Chen et al., 2005; Morgeson & Hofmann, 1999).

The goal of the current article is to explore the problems of assuming multilevel measurement and structural invariance in factor analyses, analyses of reliability, and MRCMs. More specifically, in exploring these three issues, we aim to elucidate the information lost to researchers when treating multilevel data as if they were single-level in nature, which is necessarily done when assumptions of invariance are not examined. While these issues have received some attention in the past (see Chen, Mathieu, & Bliese, 2004; Cronbach, 1976; Dyer, Hanges, & Hall, 2005; Raudenbush & Bryk, 2002; Sirotnick, 1980), such discussions seem to have gone unrecognized by researchers. Additionally, these discussions typically explore multilevel analyses, such as reliability (e.g., Chen et al., 2004), from the perspective of construct validity, rather than assumptions of cross-level invariance, and have failed to provide an integrative exploration centered on assumptions of cross-level measurement *and* structural invariance. Below, the nature of these assumptions, and the effects of their being violated, is explored. We couch our discussion of these assumptions in terms of how each relates to the separation of within- and between-groups variance—termed "disaggregation" (Muthén & Satorra, 1995)—and the meaning of these sources of variation for data collected at a single level of analysis. In addition to offering this conceptual discussion, we also provide empirical examples and potential solutions to these issues using observed data. We begin with a general exploration of multilevel data and their analysis.

Multilevel Data and Their Analysis

Multilevel data are those that contain sources of variance at multiple levels of analysis (Muthén & Satorra, 1995). A typical example of such data is that gathered from individuals nested within groups. When assessing individuals within such groups along a measure of interest, the total variance in the responses will be constituted by both between- and within-groups variance. So, data collected from individuals within groups contains (at least) two identifiable sources of variance: between- and within-groups variance. Between-groups variance is the variance that exists between the groups, while within-groups variance is the variance that exists within the groups (Kenny & La Voie, 1985). This straightforward concept represents the proverbial backbone of, and impetus for, multilevel modeling.

Each group's average value along the measure will be that group's contribution to between-groups variance. This is because the mean value for the group cannot vary within the group, only between the groups. That is, at the group-level of analysis, the only value which can vary—given that all individual-level scores are equally weight-

ed—is the group's mean. Similarly, each individual's deviation from their respective group mean is his or her contribution to within-groups variance. Again, this is because the groups' means cannot vary within the groups; only deviations away from a group's mean can vary within a group. Thus, at the within-groups level of analysis, the only value which can vary is the deviation away from the group's mean.

With multilevel data, the between-groups variance may be thought of as "group-level" variance because it is independent of the within-groups variance. Similarly, deviations away from the group means may be thought of as "individual-level" variance because it is independent of the group effect. This logic of between- and within-variance, as applied to an observed variable, may be represented as:

$$y_{ij} = y_j + (y_{ij} - y_j) \quad (1)$$

where y_{ij} is a score for a person i in a group j and y_j is the average value for group j. As can be read, y_j is the group's contribution to between-groups variance. Conversely, $(y_{ij} - y_j)$ is the individual's contribution to within-groups variance. Importantly, these "disaggregates" are additive and orthogonal (Cronbach & Webb, 1975). In other words, (a) by summing these two values, the original score along the measure may be obtained and (b) separate, independent structures may be posited as influencing these two sources of variation, meaning that we may separately model relationships at the between- and within-groups level of analysis (Muthén, 1989, 1991, 1994, 1997; Rabe-Hesketh, Skrondal, & Pickles, 2004; Skrondal & Rabe-Hesketh, 2004).

Now, consider a case in which the observed variable y_{ij} is one of many items within a scale, which may be represented as y_{pij}, where p represents the item with which the datum is associated. Given meaningful amounts of variance associated with both j and i, these items will contain both between- and within-groups variance, respectively. As such, any parameter estimates describing these items' functioning which are insensitive to this multilevel variance will reflect both sources of variation and will constrain any summaries of the items' functioning to equality across the levels of analysis. Similarly, any single parameter which describes these items' relation to other variables will reflect both between- and within-groups effects—provided the other variables also contain both sources of variation—thereby constraining any effect estimates to equality across the levels of analysis (see Raudenbush & Bryk, 2002). In other words, the multilevel sources of variation in these data will be ignored and assumptions of multilevel invariance will be imposed if one uses single-level analyses. Below we demonstrate how this effect occurs for both single-level reliability and confirmatory/exploratory factor analyses. In addition, we also show how this effect occurs in analyses in which researchers recognize and explicitly model their multilevel data.

Data

To provide an example of these issues for the current work, we used a dataset gathered from a large Northeastern organization whose focus is largely one of customer service (see Tables 1 and 2 for descriptive statistics). This organization has a structure which allows individuals to be nested into customer-service oriented groupings across the country, with 6,572 individuals nested within 505 groupings. The data were collected as part of an annual company-wide survey which took place in 2001. All items used for the current study were answered along a Likert-type scale with five response options, which ranged from strongly disagree to strongly agree—more details regarding the variables are provided below.

Confirmatory and Exploratory Factor Analysis

Confirmatory and exploratory factor analyses are commonly conducted to assess the functioning of items within a scale. These analyses give insight into latent influences which may be reflected across any measures of interest; "[they are] a method for *classification of linear dependence*" (Jöreskog, 1979, p.10, original emphasis). Ergo, factor analyses may be conceptualized as extracting sources of covariation among a series of observed variables, sources which allow for explaining observed variables with a smaller number of latent variables (i.e., factors; Gorsuch, 1983).

We may represent the model-implied covariance structure associated with the common factor model as

$$\Sigma = \Lambda\Psi\Lambda' + \Theta \quad (2)$$

where Σ is a population variance/covariance matrix of observed variables, Λ is a matrix of factor loadings (and Λ' is its inverse), Ψ is an identity matrix of factor variances, and Θ is an identity matrix of residual variances. With such a structure, the variables which constitute Σ may be assumed linearly independent conditional on their relationships with the latent variables.

However, now assume that Σ contains both between and within-groups sources of variance/covariance (i.e., data have been gathered from people in groups). The above factor model may be reconstructed to show the implications of such multilevel variance/covariance as

$$\Sigma_T = \Lambda_T \Psi_T \Lambda_T' + \Theta_T \quad (3)$$

where

$$\Sigma_T = \Sigma_B + \Sigma_W \quad (4)$$

$$\Lambda_T = \Lambda_B + \Lambda_W \quad (5)$$

$$\Psi_T = \Psi_B + \Psi_W \quad (6)$$

$$\Theta_T = \Theta_B + \Theta_W \quad (7)$$

and, inserting the elements in Equations 4–7 into Equation 3, it follows that

$$\Sigma_T = \Sigma_B + \Sigma_W = (\Lambda_B + \Lambda_W)(\Psi_B + \Psi_W)(\Lambda_B' + \Lambda_W') + (\Theta_B + \Theta_W) \quad (8)$$

where all terms have the same meaning as in Equation 2, but the subscript *T* represents a total, aggregate matrix, and the subscripts *B* and *W* represent between- and within-groups matrices, respectively. With multilevel data, Equations 3 and 8 show how constraints of invariance are placed on the common CFA/EFA across the between- and within-groups parts. There are a number of problems with such constraints.

First, the CFA/EFA solution constrains to invariant the factor structures across the levels of analysis, meaning the possibility of different numbers of factors across levels is not allowed because Ψ_T may take only one form. Again, this is a problem not only because such an assumption may be unjustified, but, moreover, because discovering differential factor structures across levels of analysis may be of substantive interest to researchers (see Cronbach, 1976; Sirotnik, 1980). Second, Equations 3 and 8 force to invariant the factor loadings and residual variances across the levels of analysis. This is of issue because such invariance is an empirical question, and while some authors argue for the justification of collapsing across multiple levels of analysis when conducting CFA/EFA (e.g., Myers et al., 2004), the final word on such invariance must always be provided by a researcher's theory and data.

Beyond the issues of parameter invariance, another potential problem associated with conducting single-level CFAs/EFAs with multilevel data is the fact that the sample sizes for the *B* and *W* parts will never be equivalent; there will always be more people than groups. This means that the parameters at the between-groups level should be assumed to be more unstable than those at the within-groups level of analysis, and when inte-

Table 1
Single-Level Means, Standard Deviations, and Correlations Among Study Variables

Item	*M*	*SD*	*ICC*	S1	S2	S3	S4	S5	S6	S7	B	AS
S1	2.59	.98	.05									
S2	2.77	1.10	.07	.53	—							
S3	2.48	1.01	.04	.45	.58	—						
S4	2.26	.94	.03	.38	.40	.44	—					
S5	2.80	1.16	.07	.54	.57	.46	.39	—				
S6	2.59	.96	.05	.45	.69	.60	.42	.55	—			
S7	2.54	.99	.06	.50	.50	.54	.42	.52	.53	—		
B	2.69	.97	.02	−.14	−.10	−.06	−.08	−.12	−.11	−.11	—	
AS	2.57	.77	.08	.73	.82	.77	.64	.77	.80	.76	−.14	—

Note. ICC = intra-class correlation, equal to the between-groups variance divided by total variance for a given variable, is an "ICC(1)"; S1–S7 = satisfaction items 1–7; B = bureaucracy; AS = average of the satisfaction items; N = 6,575.

Table 2
Disaggregated Correlations Among Study Variables

Item	S1	S2	S3	S4	S5	S6	S7	B	AS
S1	—	.78	.84	.72	.88	.78	.88	−.61	—
S2	.52	—	.89	.77	.87	.91	.76	−.22	—
S3	.44	.57	—	.82	.81	.82	.82	−.27	—
S4	.36	.39	.43	—	.73	.78	.72	−.08	—
S5	.51	.55	.44	.37	—	.82	.81	−.42	—
S6	.43	.68	.59	.40	.53	—	.79	−.31	—
S7	.48	.49	.52	.41	.50	.52	—	−.52	—
B	−.12	−.10	−.05	−.08	−.11	−.11	−.09	—	−.39
AS	—	—	—	—	—	—	—	−.13	—

Note. Lower diagonal = within-groups correlations with $N = 6575$; Upper diagonal = between-groups correlatios with $N = 505$; S1–S7 = satisfaction items 1–7; B = bureaucracy; AS = average of the satisfaction items; the correlations among the average satisfaction variable and its constituent satisfaction items was not estimated because the between-groups matrix was singular (these variable were *very* highly correlated).

grating across *B* and *W*, this should be reflected with proper weighting. However, in Equations 3 and 8, the *B* and *W* parts are given equal weight. This may be shown with an equation that has a meaning equivalent to both Equations 3 and 8

$$\Sigma_T = (1{\cdot}\Lambda_B + 1{\cdot}\Lambda_W)(1{\cdot}\Psi_B + 1{\cdot}\Psi_W)(1{\cdot}\Lambda_B' + 1{\cdot}\Lambda_W') + (1{\cdot}\Theta_B + 1{\cdot}\Theta_W) \quad (9)$$

where each matrix is given a unit weighting, leaving the final CFA/EFA solution unreflective of the fact that the *B* and *W* parts may be assumed to de differentially stable because of the differential number of observations associated with each.

This means that a final CFA/EFA solution, which integrates across the *B* and *W* parts, will be insensitive to any difference in the number of units across the levels of analysis (e.g., 50 groups and 500 individuals vs. 5 groups and 500 individuals); instead, the final solution will depend on the proportions of variance associated with each level of analysis which constitute Σ_T. For example, if Σ_B accounts for much more total variance/covariance in Σ_T than does Σ_W, Σ_B will largely drive the final factor solution (i.e., it will largely determine Λ_T, Ψ_T, and Θ_T), even though Σ_B could be associated with a much smaller sample size than Σ_W. The implication of this fact is that as the proportion of total variance associated with one level of analysis increases, that level of analysis will increasingly determine the final factor solution, irrespective of the number of observations associated with it.

All of this is to say that single-level CFAs/EFAs with multilevel data have the potential to misrepresent the underlying nature of a researcher's data (Dyer et al., 2005). However, such misrepresentations are not only undesirable, they are unnecessary. Using a logic similar to that found in Equation 1, we may disaggregate the variance in observed variables into their respective *B* and *W* parts, and conduct CFAs or EFAs at each level of analysis. To do this, however, requires estimating the population variance/covariance matrices Σ_B and Σ_W, which are estimated with S^*_B and S_{PW}, respectively. Here, S_{PW} is the pooled covariance matrix of the within-group scores and S^*_B is the covariance matrix of the between-groups scores, scaled for group size (Muthén & Satorra, 1995). These are

$$S_{PW} = \frac{\sum_j^k \sum_i^n (y_{ij} - y_j)(y_{ij} - y_j)'}{N - k} \quad (10)$$

which is a variance/covariance matrix of deviations for each individual away from each *j*'s group mean across all *n* individuals within each group and across all *k* groups (Muthén, 1994), and

$$S^*_B = \frac{\sum_j^k n(y_j - y)(y_j - y)'}{k - 1} \quad (11)$$

which is the variance/covariance matrix of means of *y* for each group *j*, weighted by *n*

(Muthén, 1994). Above, y_{ij} and y_i are as previously defined, y is the grand mean of y across all i and j units, k is the total number of groups, n is the number of individuals within each group, and N is the total sample size—to compute the B and W correlation matrices, one need only divide the covariances by the product of the standard deviations of the variables at their respective level of analysis. While S_{PW} is a consistent estimator of Σ_W, S^*_B contains elements of both Σ_W and $c\Sigma_B$, where c is a scaling factor representing the common group size (see Muthén, 1991). Therefore, CFAs and EFAs using S_{PW} may be interpreted simply, but those using S^*_B should be considered as slightly more exploratory than traditional factor analyses (Muthén, 1991).

To highlight the possible differences found when conducting single-level and multilevel factor analyses with multilevel data, we conducted both types of analyses using an employee job satisfaction scale with seven items, which addressed satisfaction with training availability, advancement opportunities, and the like (see Tables 1 and 2 for descriptive statistics). These items were self-referential in that each item assessed the satisfaction of the individual. Thus, the between-groups model of these variables was the variance/covariance of group-averaged job satisfaction and the within-groups model of these variables was the variance/covariance of deviations away from group-averaged job satisfaction—although the satisfaction items are self-referential, the between-groups variance in satisfaction is relevant given that various group- and organization-level factors have been shown to be associated with satisfaction (see Howard & Frink, 1996).

Using the total, aggregate correlation matrix, a single-level EFA was conducted using a maximum likelihood estimator with promax rotation with the program Mplus (see Muthén & Muthén, 1998–2006). As is shown in Model A from Table 3, a 3-factor solution fit the data the best. Looking at the factor loadings shown as Model A from Table 4, and using a cutoff loading of .40 and above (in concert with smaller loadings on all other factors) to assign variables to factors, it is clear that items 1 and 5 load on factor 1, items 2 and 6 load on factor 2, and items 3, 4, and 7 load on factor 3.

However, when we disaggregate the correlation matrix as allowed in Equations 10 and 11, this is no longer the case. Again using the Mplus program, both S_{PW} and S^*_B were computed (see Table 2) and an EFA was conducted on each matrix for the 7-item satisfaction scale. While the overall number of factors was stable across the B and W parts (see Models B and C from Table 3), the factor loadings and concomitant factor structures were discrepant across the levels of analysis. As shown in Model B from Table 4, the structure of the single-level EFA is maintained at the within-groups level of analysis (due to the much larger proportion of variance associated with the within-groups loadings, as indicated by the intraclass correlation coefficients found in Table 1).

Problematically, for interpreting the single-level EFA, the single-level factor structure does not match the results found at the between-groups level of analysis. As is shown in Model C from Table 4, items 1, 5, and 7 load on the first factor, only item 2 relates to the second factor, and the third factor is now composed of items 3, 4, and 6. In other words, the pattern of factor loadings are noninvariant across levels of analyses. This type of noninvariance has been called "configural noninvariance" in literature on invariance testing (e.g., Vandenberg & Lance, 2000). This noninvariance means that the latent variables cannot be assumed

Table 3
Results of Factor Analyses with Promax Rotation

	1 factor				2 factors				3 factors			
	χ^2	*RMSEA*	*SRMR*	*df*	χ^2	*RMSEA*	*SRMR*	*df*	χ^2	*RMSEA*	*SRMR*	*df*
Model A	838.463	.095	.038	14	168.923	.200	.029	8	70.366	.058	.010	3
Model B	506.944	.264	.037	14	300.505	.075	.023	8	31.329	.137	.009	3
Model C	801.070	.093	.038	14	372.482	.083	.026	8	69.293	.058	.010	3

Note. Model A = single-level exploratory factor analysis; Model B = between-groups exploratory factor analysis; Model C = within-groups exploratory factor analysis; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; *df* = degrees of freedom.

Table 4
Promax-Rotated Factor Loadings and Factor Correlations

	Single-level			Between-groups			Within-groups		
Item	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
S1	.673	.032	.082	.913	.005	.067	.676	.021	.084
S2	.241	.680	.044	.082	.998	.103	.250	.669	.031
S3	−.020	.223	.615	.324	.164	.538	−.029	.238	.594
S4	.134	.004	.472	.083	−.031	.851	.121	.003	.475
S5	.480	.178	.164	.702	.162	.134	.446	.175	.183
S6	−.006	.529	.386	.248	.257	.496	−.023	.560	.362
S7	.269	−.043	.563	.667	−.054	.323	.229	−.032	.580
F1	—			—			—		
F2	.571	—		.634	—		.567	—	
F3	.696	.647	—	.778	.684	—	.689	.649	—

Note. S1–S7 = satisfaction items 1–7; F1–F3 = factors 1–3 for indicating factor correlations.

to have equal meanings across the levels of analysis; the items are reflecting different latent variables across the levels of analysis.

As can be seen, these results indicate the benefits of taking a multilevel approach to measurement-based issues with multilevel data. Again, as is shown Table 4, an assumption of multilevel measurement invariance may be imposed in single-level EFAs of multilevel data. In this case, the assumption of multilevel measurement invariance—of the strict-invariance type forced by the single-level factor analysis—is untenable, given the difference in the pattern of factor loadings across the levels of analysis. From a theoretical standpoint, when finding such noninvariance across levels of analysis, a researcher should attempt to interpret the meaning of the latent variables across the levels of analysis based on the substantive content of the items in question and on the context within which they were collected (see Cronbach, 1976; Sirotnik, 1980). Based on theoretical considerations, one now might conduct a series of multilevel CFAs, for instance, to better understand the meaning of the latent variable and the causes of these discrepant loadings (see Dyer et al., 2005).

Finally, the issue of discrepant factor structures aside, another notable result of the multilevel EFA was the much larger factor loadings found at the between-groups level of analysis, indicating that the between-groups variance may be considered much more reliable than the within-groups variance. Regarding this difference, it is first important to note that examining the magnitude of these differences across levels of analysis with tests of statistical significance is problematic and, as with many issues of interpretation in the context of EFA results (e.g., number of factors to retain; for a review, see Gorsuch, 1983), researchers should interpret any differences in light of their theoretical model and observed data. In particular, attention should be given to the pattern of similarities and differences across levels, not on particular, isolated discrepancies among parameter estimates. The former approach provides for a more theoretically sound explanation of results and also prevents assigning too much weight to a particular finding that potentially represents an anomaly within the sample data.

Also, when attempting to meaningfully interpret differences in factor loadings across levels researchers should be aware that, because the between-groups part of the model occurs as a function of aggregates (i.e., group means), much of the measurement error in the data is removed in the between-groups model (see Cronbach, 1976). This fact means that researchers may expect larger between-groups factor loadings in general—albeit less stable factor loadings in the between-groups model owing to the smaller sample size between groups than within groups. Therefore, researchers may find most useful any differences in the pattern of factor loadings across levels of analysis (i.e., configural invariance across levels) rather than the magnitude of the difference between any given factor loading across levels. To address the difference in reliabilities across levels, we now explore single-level reliability indices computed with multilevel data.

Reliability Analysis

A prerequisite for any psychological test having meaning or practical utility is the test being a reliable measure of the construct or attribute of interest. Nunnally (1978, p. 206) defined reliability as "the extent to which [measurements] are repeatable and that any random influence which tends to make measurements different from occasion to occasion is a source of measurement error." Analytically, reliability represents the proportion of covariance, or systematic variance, relative to the total variance of a measure (Lord & Novick, 1968). Accordingly, a set of measurements exhibits reliability to the extent that it is composed of systematic variance and is free from random-error variance. Whereas tests exhibiting "adequate" reliability may or may not measure the intended construct, tests lacking reliability represent no more than random variance and, therefore, necessarily do not reflect the underlying attribute (or, for that matter, any one attribute). Related to this point, reliability is important to assess because it places an upper-bound limit on relationships between a scale score composed of a group of items and any other variables (Cortina, 1993).

Problematically, when reliability is computed for a scale which was administered to individuals within groups, the reliability value will reflect both between- and within-groups sources of consistent variation (Muthén, 1991). In other words, this value will collapse across between- and within-groups variation and constrain to invariant each source of internal consistency. For example, consider a scale in which the lower-level dependent variables contain both between- and within-groups variance (e.g., 5% between-groups and 95% within-groups variance). Perhaps the scale is completely unreliable at the between-groups level, but is perfectly reliable at the within-groups level (an unlikely event [Bliese, 1998], but discussed here for pedagogical purposes). In such a case, a single-level reliability computation will surely return an acceptable value because only 5% of the variance is unreliable. However, the relationship between a group-level predictor and this lower-level dependent variable necessarily would be zero, due to the complete unreliability at the between-groups level. Further, the relationship between this variable and any other variable which contains between-groups variance also will be restricted to zero at the between-groups level—something a researcher might erroneously conclude had to do with a lack of a substantive relationship, rather than a measurement-based problem, because of the adequate single-level reliability value which was observed.

Conversely, consider the case in which the same dependent variable is perfectly reliable at the between-groups level but not at the within-groups level. By computing alpha, the researcher may conclude that the measure is extremely unreliable (as 95% of the variance is not internally consistent), failing to consider the fact that the measure is perfectly reliable at the between-groups level (a much more likely event [Bliese, 1998]). In this case, again, simply computing single-level reliability would yield misleading or erroneous results in terms of the measure's properties, the meaning of the variables, and possible predictor-criterion relationships. Further, by only examining single-level reliability, a researcher would be unaware of the fact that the between-groups portion of the scale mean would be able to have nonzero relationships with other variables, while the within-groups portion of the scale mean would not.

As noted above, the erroneous assumption of invariance in the above examples is problematic not simply because the invariance constraint is empirically unjustified. Instead, in both cases the researcher has missed the opportunity to discover important information regarding the functioning of the observed variables. With such information, the researcher would be able to change not only their interpretation of the observed variables from a theoretical perceptive (see Chen et al., 2004) but, concurrently, change the analytic tact taken in the research endeavor.

In order to contrast single-level and multilevel reliability estimates found with the satisfaction scale above, we first computed single-level alpha for the satisfaction items. This yielded a value of .88. However, recalling above that the between-groups variance in the items had much higher factor loadings on the latent variables, we can expect the value of .88 to be collapsing across the between-groups reliability, which should be higher than .88, and the within-groups reliability, which should be lower than .88. To investigate this proposition, we disaggregated the between- and within-groups variance in the items in line with Equation 1. In other words, we computed

each group's mean for the satisfaction items and each individual's deviation away from the group mean. Then we computed alpha for the between-groups data, which was .93. This stands in contrast to the within-groups alpha value, which was .87.

As is clear with our data, the reliabilities are different—although they are both adequate—and the within-groups variance has largely determined the reliability estimate. This has implications for MRCMs which would use a mean of these satisfaction items. Aside from the issues mentioned above, one important aspect of the discrepant reliabilities has to do with corrections researchers may desire to make to relationships associated with scale aggregates. As is well known, one may correct relationships for any attenuation due to unreliability using Spearman's classic formula (see Nunnally, 1978). However, if one were to use the single-level reliability estimate, any such corrections would not take into account the fact that the correction at the between-groups level should be smaller than the correction at the within-groups level. We now explore assumptions of structural invariance which are often committed with multilevel data.

Multilevel Random Coefficient Modeling

As noted above, MRCM allows researchers to answer questions at the level at which they are asked (Nezlek & Zyzniewski, 1998). This is because the MRCM framework can disaggregate a variable's variance into the appropriate between- and within-groups components. This may can be illustrated in the structure of the common MRCM (see Raudenbush & Bryk, 2002), which may be heuristically represented as

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + r_{ij} \quad (12)$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (13)$$

$$\beta_{1j} = \gamma_{10} \quad (14)$$

where y_{ij} is as above, β_{0j} is the intercept for a group j (i.e., the group mean in a model without predictors), β_{1j} represents the effect of a predictor x on the criterion y for a group j, x_{ij} is the score along a predictor x for person i in group j (a predictor which contains both between- and within-groups variance), r_{ij} is the residual in the model for person i in group j, γ_{00} is the grand-intercept (i.e., the grand mean in a model without predictors), u_{0j} is the intercept-residual for group j (i.e., the group's distance from γ_{00}, which allows the group's intercept to randomly vary around γ_{00}), γ_{10} is the grand slope-intercept (i.e., the average relationship between x and y for the model) – models with a randomly varying β_{1j} (i.e., u_{1j}) are not discussed here for the sake of concision.

As may be surmised, the variance across individuals in r_{ij} is within-groups variance, while the variance across groups in u_{0j} is between-groups variance. In other words, this model has disaggregated "residual" variance (i.e., variance unaccounted for by any substantive predictors) in y_{ij} at the appropriate level of analysis. However, less easily surmised is the fact that the between- and within-groups variance in this model may not only be viewed as "residual" variance but may also be conceptualized as being associated with the effect of the predictor on the criterion (it may be helpful to recall that, in this example, *all* of the variance in y_{ij} is either between- or within-groups variance, not just between- or within-groups residual variance). This is because the terms β_{1j} and γ_{10}, the latter of which is often reported in multilevel analyses (e.g., Hofmann et al., 2003), collapse across both the between- and within-groups influences of x on y, effectively constraining these structural parameters to invariance across multiple levels of analysis.

The constraint of structural invariance in MRCMs may be illustrated by using the mean of the satisfaction items mentioned above (i.e., the scale mean for each person), and regressing that mean onto a variable which assesses employees' perceptions of the level of bureaucracy up with which they have to put (again, see Tables 1 and 2 for descriptive statistics). Drawing from the job characteristics model (see Hackman & Oldham, 1976), it is reasonable to expect a relationship among these variables to the extent that less bureaucracy in one's job should allow increased individual performance, leading to greater levels of "experienced meaningfulness", and therefore greater levels of job satisfaction. Using the satisfaction-scale mean as the criterion and the bureaucracy variable as a predictor, we may estimate the model shown in Equations 12–14 (it is noted that first an "unconditional model", using only the satisfaction variable, was estimated for continuity [see

Table 5
Multilevel Random Coefficient Model of Multilevel Effects

Effect	G	StdG	*SE*	*T*	*df*	*p*
Model A						
INTERCEPT, γ_{00}	2.564	—	.012	187.119	504	<.05
Model B						
INTERCEPT, γ_{00}	2.541	—	.012	189.230	504	<.05
INTERCEPT, γ_{10}	−.102	−.133	.004	−8.697	504	<.05
Model C						
INTERCEPT, γ_{00}	4.236	—	.587	7.095	504	<.05
SLOPE, γ_{01}	−.667	−.390	.241	−2.768	504	<.05
INTERCEPT, γ_{10}	−.098	−.128	.012	−8.298	504	<.05
	Parameter variance	*SE*	*t*	*df*	*p*	
Model A						
Between-groups variance, u_{0j}	.046	.006	7.461	504	<.05	
Within-groups variance effect, r_{ij}	.548	.012	46.948	504	<.05	
Model B						
Between-groups variance, u_{0j}	.044	.006	7.298	504	<.05	
Within-groups variance effect, r_{ij}	.541	.012	44.574	504	<.05	
Model C						
Between-groups variance, u_{0j}	.039	.007	5.845	504	<.05	
Within-groups variance effect, r_{ij}	.540	.012	44.694	504	<.05	

Note. G = γ, StdG = γ parameters standardized to the variance(s) of the variable(s); *SE* = standard error; *df* = degrees of freedom.

Table 5, Model A]). Again, this was accomplished using the program Mplus.

Modeled parameters are provided in Table 5, Model B; for estimates of "variance accounted for" at each level of analysis, comparisons may be made between initial between- and within-groups variance terms (i.e., Model A), and these terms following the inclusion of the predictor (Raudenbush & Bryk, 2002). The variance accounted for between groups was 4.35%, and the variance accounted for within groups was 1.28%. As is shown in Model B, Table 5, the term γ_{10} provides a single effect estimate for the between- and within-groups effect of x on y ($\gamma_{10} = -.102$, $SE = .012$, Standardized Parameter $= -.133$). This is an assumption of multilevel invariance because the possibility of discrepant effects across the levels of analysis has been left untested and, instead, is assumed. More specifically, this is an assumption of multilevel structural invariance, as the structural relationships between the predictor and criterion are forced to strict invariance across the between- and within-groups levels of analysis.

Interestingly, the fact that γ_{10} is a function of both between- and within-groups effects has been intimated in previous work on applying MRCM to data with a nested structure (e.g., Nezlek & Zyzniewski, 1998). Further, a simple technique exists for assuring that assumptions of multilevel structural invariance are not violated (Kreft, de Leeuw, & Aiken, 1995). By "group-mean centering" the predictor (i.e., by removing the between-groups variance in x) and entering the extracted means at the group-level of analysis, researchers may disaggregate the effect of the predictor on the criterion because there will be different effect estimates of x on y at each level of analysis. This may be shown as

$$y_{ij} = \beta_{0j} + \beta_{1j}(x_{ij} - x_j) + r_{ij} \quad (15)$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}x_j + u_{0j} \quad (16)$$

$$\beta_{1j} = \gamma_{10} \quad (17)$$

where all terms are as above (and as shown in Equation 1), such that $(x_{ij} - x_j)$ represents the within-groups portion of the predictor (i.e., deviations away from the group mean for each individual) and x_{ij} represents the between-groups portion of the predictor (i.e., the group means). Results from this model are presented in Table 5, Model C.

As is clear, this model allows for the appropriate disaggregation of not only residual variance, but also the multilevel effect of x on y, with the within-groups effect γ_{10} (= −.098, SE = .012, Standardized Parameter = −.128) and the between-groups effect γ_{01} (= −.667, SE = .241, Standardized Parameter = −.39). By comparing these disaggregated results to that derived above, it is clear that the within-groups effect was largely driving the total effect (because of the significantly larger amount of variance at the within-groups level), which served to grossly misrepresent the between-groups effect. Such misrepresentation has been solved with the logic shown in Equations 15–17. In other words, this model allows researchers to not assume multilevel structural invariance in their parameter estimates.

Further, this disaggregated model, because it allows the between-groups effect to be freely estimated, rather than be restricted to equality with the within-groups effect, allows more of both the between- and within-groups variance to be accounted for than the previous model. The variance now accounted for between groups was 16.67% (vs. 4.35%) and the variance accounted for within groups was 1.46% (vs. 1.28%). This is because the effect estimate at each level of analysis is optimized to represent the level-specific effect of x on y, and when these effects differ across levels of analysis, a single parameter representing both will not optimally solve for the equation at either level. As indicated by the difference in the between- and within-groups effects, and the proportions of variance accounted for, without disaggregating the multilevel effects in their models, researchers run the rather serious risk of inappropriately constraining structural parameters to invariance across levels of analysis; such a constraint affected not only on the effect of x on y, but also the associated "variance accounted for" statistics.

Before concluding, it is relevant to discuss the fact that Raudenbush and Bryk (2002) note that difference in effects across levels of analysis may be tested by grand-mean centering individual-level predictors (which contain both between- and within-groups variance) and entering them at "Level 1," as shown in Equations 12. Then, by entering group means of the same variable at "Level 2," as shown in Equation 16, researchers may test the difference between the within- and between-groups effects. With such a model, the effect at Level 1 becomes the within-groups effect and the effect of the group mean becomes the between-groups effect minus the within-groups effect (i.e., the composition effect [see Raudenbush & Bryk, 2002]). Thus, the statistical significance of the effect of the group means is the statistical significance of the difference across the between- and within-groups levels of analysis in the effect of the predictor on the criterion (i.e., the composition effect). Without such statistical significance, researchers may assume that the effects of interest are invariant across levels (Kenny, Bolger, & Kashy, 2001).

Discussion

In this article, we have attempted to highlight the assumptions of measurement and structural invariance that researchers often implicitly make when analyzing data that contain meaningful multilevel variance. We hasten to mention that the stated problems of assuming invariance could be far worse than the current exemplars illustrate. For example, consider a situation where a factor structure is not invariant across levels of analysis: a series of observed variables that are multidimensional at the within-groups level and unidimensional at the between-groups level of analysis, both with adequate factor loadings. Depending on the proportion of variance residing at each level and the factor analytic technique used, a factor analysis could extract either two or one factor from the observed variables. In the former case, if each factor is used to create two separate variables for use in an MRCM, the unidimensional between-groups variance will be unjustifiably partitioned into two scale scores. In the latter case, if a single variable is used to represent the items, the within-groups factors will be treated as a single variable, providing for equally problematic within-groups scale aggregates. When investigating cross-level differences in factor structures, researchers should attend to any differences, both from a theoretical and empirical perspective, and choose their subsequent analyses based on such concerns.

Moving on to assumptions of multilevel structural invariance, it should be recognized that this issue is one of both theoretical and statistical importance. Statistically, for example, as one increases (1) the number of predictors and (2) the levels of analysis which are modeled

in one's data, the possibility of inappropriate multilevel invariance becomes greater when employing techniques insensitive to the multilevel nature of effects. For example, assuming a .5 probability that effects of predictors on a criterion vary across levels of analysis for each predictor in a MRCM, a 2-level model with a single predictor will have a .5 probability of forcing inappropriate invariance on effects because there is one predictor with one difference in the level of analysis: level 1 versus level 2. However, a 3-level model with one predictor will have a .875 probability of committing this error because there are now three differences in level of analysis: level 1 versus level 2, level 1 versus level 3, and level 2 versus level 3. Now, consider the same progression, but with two predictors. The probability of an erroneous multilevel invariance assumption is .75 with a 2-level model, but for a 3-level model the probability jumps to .984375! based on the notion that complex, polylevel models are likely to become more common than they are today in multilevel research, the practice of automatically constraining relations to invariance across levels will only become more insidious with such increasingly complex models.

Due to these concerns, theoretical postulations finding support through empirical results which were, in turn, informed by inappropriate analyses of multilevel data, should be considered suspect. Also of suspicious origins are findings stemming from analyses which have left unexamined the possibility of noninvariant cross-level measurement and structural parameters estimates. Regarding the former, any research which explores the factor structure of multilevel data using single-level analyses may be providing results which have little meaning (depending on the true underlying factor structure at multiple levels of analysis). Regarding the latter, MRCM analyses which have collapsed across between- and within-groups effects, and constrained to equality these effects, should be warily used as support for hypotheses which specify relationships between "individual level" variables (when such variables contain both between- and within-groups effects). In concert, these points mean that, while multilevel research has the potential to offer new avenues of insight for theory and research of people in groups, unless it is done correctly, it has the potential to provide results that are obtuse at best, and erroneous at worst (Raudenbush & Bryk, 2002). Only research which correctly accounts for the multilevel influences on its data may offer insight into the complex phenomena which shape people and groups.

Notably, although we have discussed these issues as assumptions, one must address before proceeding with subsequent analyses, findings of invariance across levels of analysis also are of substantive value in theory and measurement development. For instance, results demonstrating that a measure of job satisfaction exhibits a different factor structure at the person and organizational level suggest the nonequivalent meaning of the job satisfaction construct at the different levels (Cronbach, 1976; Sirotnik, 1980). Such findings should engender investigations into the nature, origin, consequences, and generalilzability of these different meanings. Is it the case, for example, that individuals in different organizations construe satisfaction in discrepant ways due to their respective organizational cultures or reward systems? Such questions can be answered by testing for cross-level measurement invariance and, assuming such invariance exists, can be modeled as a function of other explanatory variables (Chen et al., 2005).

In closing, we have outlined methods for examining cross-level differences in measurement and structural parameters. Such differences carry with them information relevant to both theoretical and empirical pursuits (Morgeson & Hofmann, 1999). However, it is of note that researchers often desire to ignore the multilevel nature of their data altogether in order to test, for example, "individual-level" or group-level models (for a discussion, see Klein, Dansereau, & Hall, 1994). In individual-level modeling, researchers ignore the grouping structure of their data and estimate single-level analyses using both the between- and within-groups variance in their model. In group-level models, researchers utilize only the between-groups variance in their measures of interest. In these cases, researchers may be disinclined to investigate their data for cross-level parameter invariance but should be aware of the possibility of cross-level differences in parameter estimates for their theoretical models of interest. If the researcher deems such differences as being not of theoretical interest or if the observed cross-level differences are minimal, they may be motivated to proceed with their originally planned investigation. However, with very large differences in parameters across levels, the ability of single-level

theories and results to tell the whole story of the data may be lacking. Therefore, armed with the discussion and methods we provide for understand cross-level differences in modeled parameters, we recommend scholars investigate their data for such differences in order to better inform the theory and empirically testable assumptions upon which their research relies.

References

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

Chen, G., Bliese, P., & Mathieu, J. (2005). Conceptual framework and statistical procedures for delineating and testing multilevel theories of homology. *Organizational Research Methods, 8,* 375–409.

Chen, G., Mathieu, J. E., & Bliese, P. D. (2004). A framework for conducting multilevel construct validation. In F. J. Yammarino & F. Dansereau (Eds.), *Research in multilevel issues: Multilevel issues in organizational behavior and processes* (Vol. 3, pp. 273–303). Oxford, UK: Elsevier.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and Applications. *Journal of Applied Psychology, 78,* 98–104.

Cronbach, L. J. (with assistance of Deken, J. E., & Webb, N.) (1976). *Research on classrooms and schools: Formulation of questions, design and analysis.* Occasional Paper of the Stanford Evaluation Consortium, Stanford University.

Cronbach, L. J., & Webb, N. (1975). Between-class and within-class effects in a reported aptitude x treatment interaction: Reanalysis of a study by G. L. Anderson. *Journal of Educational Psychology, 67,* 717–724.

DeShon, R. P., Kozlowski, S. W., Schmidt, A. M., Milner, K. R., & Wiechmann, D. (2004). A multilevel, multiple goal model of feedback effects on the regulation of individual and team performance in training. *Journal of Applied Psychology, 89,* 1035–1056.

Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *Leadership Quarterly, 16,* 149–167.

Gorsuch, G. L. (1983). *Factor analysis* (2nd ed.). Englewood Cliffs, NJ: Erlbaum.

Hackman, J. R., & Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance, 16,* 250–279.

Hofmann, D. A., Morgeson, F. P., & Gerras, S. (2003). Climate as a moderator of the relationship between LMX and content specific citizenship: Safety climate as an exemplar. *Journal of Applied Psychology, 88,* 170–178.

Howard, J. I., & Frink, D. D. (1996). The effects of organizational restructure on employee satisfaction. *Group & Organization Management, 21,* 278–294.

Hox, J. (2002). *Multilevel analysis: Techniques and applications.* London, England: LEA.

Jöreskog, K. G. (1979). Basic ideas of factor and component analysis. In J. Magidson, *Advances in factor analysis and structural equation models* (pp. 5–20). New York, NY: University Press.

Kenny, D. A., Bolger, N., & Kashy, D. A. (2001). Traditional methods for estimating multilevel models. In D. S. Moskowitz & S. Hershberger (Eds.), *Modeling intraindividual variability with repeated measures data: Methods and applications* (pp. 1–24). Englewood Cliffs, NJ: Erlbaum.

Kenny, D. A., & La Voie, L. (1985). Separating individual and group effects. *Journal of Personality and Social Psychology, 48,* 339–348.

Klein, K. J., Dansereau, F., & Hall, R. J. (1994). Levels issues in theory development, data collection, and analysis. *Academy of Management Review, 19,* 195–229.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30,* 1–22.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison Wesley.

Luong, A., & Rogelberg, S. G. (2005). Meetings and more meetings: The relationship between meeting load and the daily well-being of employees. *Group Dynamics: Theory, Research and Practice, 1,* 58–67.

Morgeson, F. P., & Hofmann, D. A. (1999). The structure and function of collective constructs: Implications for multilevel research and theory development. *Academy of Management Review, 24,* 249–255.

Moritz, S. E., & Watson, C. B. (1998). Levels of analysis issues in group psychology: Using efficacy as an example of a multilevel model. *Group Dynamics: Theory, Research and Practice, 2,* 285–298.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557–585.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28,* 338–354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research, 22,* 376–398.

Muthén, B. O. (1997). Latent variable modeling of longitudinal and multilevel data. In A. Raftery

(Ed.), *Sociological Methodology* (pp. 453–480). Boston, MA: Blackwell Publishers.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 267–316). Washington DC: The American Sociological Association.

Muthén, L. K., & Muthén, B. O. (1998–2006). *Mplus user's guide: Statistical analysis with latent variables* (4th ed.). Los Angeles, CA: Muthén & Muthén.

Myers, N. D., Feltz, D. L., & Short, S. E. (2004). Collective efficacy and team performance: A longitudinal study of collegiate football teams. *Group Dynamics: Theory, Research and Practice, 8,* 126–138.

Myers, N. D., Payment, C. A., & Feltz, D. L. (2004). Reciprocal relationships between collective efficacy and team performance in women's ice hockey. *Group Dynamics: Theory, Research and Practice, 8,* 182–195.

Nezlek, J. B., & Zyzniewski, L. E. (1998). Using hierarchical linear modeling to analyze grouped data. *Group Dynamics: Theory, Research, and Practice, 2,* 313–320.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Pollack, B. N. (1998). Hierarchical linear modeling and the "unit of analysis" problem: A solution for analyzing responses of intact group members. *Group Dynamics: Theory, Research, and Practice, 2,* 299–312.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69,* 167–190.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Sirotnik, K. A. (1980). Psychometric implications of the unit-of-analysis problem (with examples from the measurement of organizational climate). *Journal of Educational Measurement, 4,* 245–282.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models.* Boca Raton, FL: Chapman & Hall/CRC.

Thoresen, C. J., Bradley, J. C., Bliese, P. D., & Thoresen, J. D. (2004). The big five personality traits and individual job performance growth trajectories in maintenance and transitional job staged. *Journal of Applied Psychology, 89,* 835–853.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–69.

Received April 22, 2006
Revision received November 14, 2006
Accepted November 14, 2006 ■