

CHAPTER 12

Multilevel Analytical Techniques

Commonalities, Differences, and Continuing Questions

Katherine J. Klein	Mark A. Griffin
Paul D. Bliese	David A. Hofmann
Steve W. J. Kozlowski	Lawrence R. James
Fred Dansereau	Francis J. Yammarino
Mark B. Gavin	Michelle C. Bligh

A variety of statistical procedures are available to establish agreement within groups and/or to analyze multilevel data. These procedures include r_{wg} , eta-squared, ICC(1), ICC(2), within-and-between analysis (WABA), hierarchical linear modeling (HLM), and cross-level operator (CLOP) analyses in regression and analysis of variance. These procedures are described in detail in Chapters Eight through Eleven. Because each of these chapters addresses a different procedure or set of procedures, readers may gain an appreciation of each procedure in isolation but may nevertheless wonder how the

Note: Preparation of this chapter was a collaborative effort. Katherine J. Klein conceived of the idea for the chapter and was the primary contact person, talking and e-mailing with the chapter authors as the chapter developed. Klein, Paul D. Bliese, and Steve W. J. Kozlowski wrote the chapter. Bliese developed the simulated data set, conducted the cross-level-operator data analyses (on the basis of Chapter Nine in this volume, by Lawrence R. James and Larry J. Williams), and conducted several of the other data analyses reported in the chapter. Michele C.

procedures compare, and whether the procedures answer the same or different questions. Indeed many debates in the literature revolve around these issues. In this chapter, we offer some answers to these questions.

We do not present an overview of each procedure. Instead, we describe commonalities and differences among the various data-analytic procedures. Accordingly, we urge readers to study Chapters One, Eight, Nine, Ten, and Eleven before reading this chapter. Our goal is to help readers make informed choices about which procedures to use when, and why; our goal is not to evaluate or advocate, nor is it to resolve differences among the procedures.

The chapter is organized into four sections. In the first section, we describe a number of dimensions that differentiate the indices often used to justify aggregation of lower-level (for example, individual) data to a higher level of analysis (for example, groups). These indices are r_{wg} , WABA, ICC(1), ICC(2), and eta-squared. In the second section, we describe dimensions that differentiate three procedures commonly used in multilevel data analysis. These procedures are WABA, HLM (one of a class of random-coefficient modeling techniques), and CLOP. In the third section, we use these indices and procedures to analyze a simulated data set, thereby illustrating similarities and differences among the approaches. In the fourth section, we offer a brief summary and concluding comments.

Aggregation of Lower-Level Data to Higher Units of Analysis

Multilevel researchers often gather data from lower-level units (for example, individuals) to assess higher-level (for example, group) constructs. In doing so, multilevel researchers typically propose, implicitly or explicitly, a composition model linking the individual-level data to the group-level construct they seek to measure (see Chan,

Bligh conducted the WABA analyses under the guidance of Fred Dansereau and Francis J. Yammarino. Mark B. Gavin conducted the HLM data analyses in consultation with David A. Hofmann and Mark A. Griffin. This chapter is based in part on Chapters Eight, Nine, Ten, and Eleven, whose authors had the opportunity to provide comments on and sign off on this chapter. Although supportive of this chapter, Williams was unavailable to provide detailed comments and therefore opted not to be included as a chapter author.

1998; Rousseau, 1985; see also Kozlowski & Klein, Chapter One, this volume). Composition models of this type propose that the lower-level data are sufficiently homogeneous within units that the data may be meaningfully aggregated to the unit level. The lower-level data are thus "candidates" for aggregation to the higher level. A number of procedures are available to determine whether these candidates are "contenders"—that is, whether the lower-level data are indeed sufficiently homogeneous to justify aggregation. These procedures include r_{wg} , WABA, ICC(1), ICC(2), and eta-squared. We distinguish between WABA and eta-squared because, although WABA utilizes eta-squared values in drawing inferences, not all multilevel researchers interpret eta-squared values in the same way that WABA researchers do. Here, we highlight five dimensions that differentiate the various procedures most commonly used to justify aggregation.

1. *Does the procedure assess variability within one unit at a time, or does the procedure assess within- and between-unit variability across the entire sample of units?*

Of the five procedures commonly used to justify aggregation of lower-level data to the higher-level of analysis, only r_{wg} (James, Demaree, & Wolf, 1984) provides an assessment of agreement within a specific group. The r_{wg} procedure compares the observed within-group variability to the within-unit variability expected from a hypothetical random distribution—that is, an expected variance, or EV. Thus a researcher uses r_{wg} to obtain a separate assessment of within-unit agreement on the measure of interest for each unit in his or her sample. Kozlowski and Klein (Chapter One, this volume) call this a *construct-by-unit approach*. This approach allows one to test whether within-group agreement is uniform across all the units in the sample.

In contrast, the remaining procedures—ICC(1), ICC(2), eta-squared, and WABA I—assess within-unit homogeneity by comparing between-group variance to the total variance across a researcher's entire sample of units. These procedures provide an omnibus index of homogeneity and do not allow one to assess whether a specific unit shows greater than chance homogeneity. Kozlowski and Klein (Chapter One, this volume) call this a *construct-by-sample*

approach. Using this approach, a researcher may conclude that a variable is homogeneous within units across the entire sample even if some specific units show little within-unit homogeneity.

Different procedures used to justify aggregation may yield different conclusions. If agreement within each unit is high but the means for the units vary little from unit to unit, then a researcher using r_{wg} alone will conclude that aggregation to the unit level is appropriate. (If so, the researcher may discover, in testing group-level hypotheses, that the sample lacks sufficient between-unit variability to provide an adequate test of the hypotheses.) In contrast, a researcher using ICC(1), ICC(2), eta-squared, or WABA may conclude that aggregation is not appropriate, given the lack of between-group variability. Conversely, if the means for the units vary considerably across a sample but variability within each unit is also relatively high, then researchers using ICC(1), ICC(2), eta-squared, or WABA may conclude that aggregation is warranted, whereas a researcher using r_{wg} may conclude that aggregation is not warranted. This potential for disagreement stems, in large part, from how random variance is defined: either in terms of an expected random distribution within a group or in terms of total (within- and between-group) variance. For more detail, see Bliese (Chapter Eight, this volume) and Dansereau and Yammarino (Chapter Ten, this volume).

2. *Is the procedure commonly used to decide between two patterns of variability (either homogeneous groups or independent individuals), or is the procedure commonly used to decide among three patterns of variability (homogeneous groups, independent individuals, or heterogeneous interdependent individuals within groups)?*

In using r_{wg} , ICC(1), ICC(2), and/or eta-squared to determine whether it is appropriate to aggregate an individual-level variable to the unit level, most researchers consider just two possibilities:

1. Yes, the variable shows sufficient within-group agreement and/or sufficient between-group variability to justify aggregating the variable to the unit level; the variable is indeed a measure of a unit property.
2. No, the variable does not show sufficient within-group agreement and/or sufficient between-group variability to justify aggregation; the variable is a measure of individual differences.

There is, however, a third possibility: specifically, a variable may show high variability within groups and low variability between groups (that is, greater within-group variability than expected by chance), which is potentially indicative of a heterogeneous or frog-pond property. This possibility, potentially overlooked by researchers using r_{wg} , ICC(1), ICC(2), and eta-squared, plays an integral role in WABA (Dansereau & Yammarino, Chapter Ten, this volume). Researchers using other procedures might dismiss substantial within-unit variability as between-individual variability or as error rather than seeing it as evidence for meaningful interdependence of individuals within groups (parts).

Note, however, that this focus on the parts condition is based more on convention than on mathematical limitations of the other approaches. For example, evidence of the parts or frog-pond condition is provided by a negative ICC(1) value, when ICC(1) is calculated from a one-way random-effects ANOVA (see Bartko, 1976; Bliese, Chapter Eight, this volume). Similarly, r_{wg} will also yield negative values potentially indicative of a frog-pond condition when within-group variance exceeds the expected variance (Brown, Kozlowski, & Hattrup, 1996).

3. *Does the procedure consider each variable separately, or does it consider the entire set of hypothesized independent and dependent variables?*

Researchers using WABA typically examine within- and between-group variability in all their measures before deciding whether to analyze the relationships among all their measures at a single level of analysis—that is, at the unit-level (wholes), within units (parts), or within and between units (equivocal or individual-level analyses). In contrast, researchers using r_{wg} , eta-squared, ICC(1), and/or ICC(2) focus on what these indices reveal about a specific variable in isolation, without necessarily considering implications for other variables in the data set. As we explain in more detail in the following section, this difference between WABA and the other aggregation indices reflects the fact that WABA is primarily designed to identify the key level or levels at which a set of relationships should be analyzed. WABA researchers seek to learn at what level or levels an effect—a relationship—occurs. Finding evidence that a relationship occurs at two different levels of analy-

sis, WABA researchers would typically analyze the relationship at the two different levels of analysis but would do so sequentially—in two separate equations.

Other analytical tools—most notably hierarchical linear modeling and cross-level operator analyses—allow researchers to test models that simultaneously incorporate variables at multiple levels of analysis. Before testing these cross-level models (see Kozlowski & Klein, Chapter One, this volume), researchers may use r_{wg} , eta-squared, ICC(1), and/or ICC(2) to justify aggregating only the specific measures that presumably reflect higher-level constructs.

4. *How and where does the procedure draw the line in determining whether the data should or should not be aggregated to the unit level?*

Researchers have long debated—and seem likely to continue to debate—the precise standards to use in deciding whether aggregation to the unit level is or is not justifiable. A common rule of thumb is that aggregation is justifiable if the r_{wg} for a unit is .70 or higher. If one is using r_{wg} to test the appropriateness of aggregation, should one remove variables and/or units with r_{wg} values that fall below .70? The answer is open to debate. Many researchers who use r_{wg} simply report the average r_{wg} for their groups (for each measure they wish to aggregate to the group level), using average r_{wg} values near or above .70 to justify aggregation. Reporting the range of r_{wg} values and the percentage of units with values greater than .70 seems reasonable. Often overlooked, however, is the importance of testing the statistical significance of r_{wg} , given various assumptions about its distribution (see Bliese, Chapter Eight, this volume).

Researchers who use eta-squared or ICC(1) to justify aggregation typically assert that aggregation is warranted if the F-test is statistically significant. A statistically significant F-test indicates that the between-group variance of a measure is significantly greater than the within-group variance of the measure. Note that the same F-test is used to evaluate the statistical significance of both eta-squared and ICC(1).

WABA researchers argue that aggregation of a measure to the unit level is justified if the F-test for the measure is statistically significant and if between-group variability in the measure is also

"practically significant." WABA researchers calculate the E-test to evaluate practical significance (see Dansereau and Yammarino, Chapter Ten, this volume). Their goal in assessing practical significance is to identify effects that are expected to hold regardless of sample size.

Yet another perspective is that aggregation is appropriate if the group means of the aggregate variable can be reliably differentiated (see Bliese, Chapter Eight, this volume). ICC(2) provides such a measure. One can interpret the ICC(2) as one would any other reliability measure. Common practice suggests that values equal to or above .70 are acceptable, values between .50 and .70 are marginal, and values lower than .50 are poor. Note that small between-unit differences, as evidenced by relatively small ICC(1) values, may lead to reliable mean differences if the units are relatively large.

In summary, the existing literature provides few hard-and-fast (and several varying) standards to use to determine whether a variable has an acceptable amount of homogeneity to justify aggregation. In many cases, the differing procedures and criteria used to justify aggregation may yield similar conclusions. Thus r_{wg} , ICC(1), ICC(2), eta-squared, and WABA may all suggest that the same variable(s) should be aggregated, but this is not always the case, as we show in our analyses of the simulated data set in the third section of this chapter.

5. *To what extent are the values obtained in using the procedure influenced by the size of the groups in the sample?*

Clearly, this is an important question for evaluating the quality of any statistical procedure used to justify data aggregation. Indeed, this question was one of the important issues underlying the debate between Schmidt and Hunter (1989) and Kozlowski and Hattrup (1992) regarding r_{wg} . Kozlowski and Hattrup (1992) demonstrated that the method for indexing within-group homogeneity recommended by Schmidt and Hunter (1989)—the standard error of the mean and its associated confidence interval—was biased by sample size. The larger the size of the groups, the more likely Schmidt and Hunter's approach was to indicate within-group homogeneity, even when actual within-group agreement was low. In contrast, r_{wg} was relatively unaffected by group size. However, r_{wg} was somewhat

attenuated when group size was ten or smaller and when actual agreement was moderate to low.

The influence of group size also lies at the heart of a quite recent and by no means resolved controversy in the levels-related literature, but this one involves bias due to small group size. WABA relies on the calculation of the E-ratio (based on eta-squared) to determine the practical significance of within- and between-unit variability in an individual-level measure. WABA advocates argue that the E-ratio is valuable because the practical significance of the E-ratio does not vary as a function of sample size (see Dansereau & Yammarino, Chapter Ten this volume). In contrast, the statistical significance of the F-ratio—considered, in WABA, in conjunction with the E-ratio—does vary as a function of sample size.

Bliese and Halverson (1998) have argued, however, that eta-squared (on which the E-ratio rests) varies as a function of sample—specifically, unit size. Bliese (Chapter Eight, this volume) suggests that small units typically yield larger eta-squared values than do large units, and that therefore a sample of two hundred individuals grouped in small units (for example, two-person dyads) is more likely to yield a practically significant E-ratio than is a sample of the same number of individuals grouped into larger units (say, twenty-person teams). Bliese proposes that ICC(1) provides a measure of between-unit variability that is not influenced (biased) by the size of the units in one's sample. Dansereau and Yammarino (Chapter Ten, this volume) counter that such results will occur in random data but not necessarily in field data, and that ICC(1) values do not vary as a function of group size precisely because the formula for ICC(1) includes and thus controls for degrees of freedom.

We have no resolution to offer. Suffice it to say that additional research is needed regarding the influence of sample size (that is, the number and size of the units in a sample) on the indices—particularly ICC(1), eta-squared, and the E-test—used in justifying aggregation to the unit level.

To summarize, we know that, within the limits already noted, r_{wg} values are generally not influenced by the size of the units in a sample (see Kozlowski & Hattrup, 1992, for details). Further, we know that ICC(2) and the significance of the F-test for ICC(1) and eta-squared are influenced by the number and size of the units in a sample. But experts, for now at least, disagree about the influence

that the number and size of the units in a sample have on eta-squared and the E-test used in WABA.

Summary Comments

Table 12.1 summarizes our comparison of r_{wg} , eta-squared, ICC(1), ICC(2) and WABA. We urge researchers to use a number of these indices before aggregating their measures. A researcher's theoretical framework may suggest that one or more of the indices is most appropriate. The different indices that we have compared are most likely to lead to the same conclusion in relatively unambiguous cases—that is, when a measure's variability lies quite predominantly between units, or when a measure's variability lies quite predominantly within units, or when a measure shows random variability within and between units. When the indices lead to differing conclusions regarding the merits of aggregation, researchers may rely on theory, prior research, and/or belief in the superior merits of one of the indices in deciding whether to aggregate their measure(s). Our analyses of the simulated data set allow us to compare the conclusions that a researcher would reach in using differing indices to justify aggregation. But we have analyzed just one data set, a simulated one at that. Additional research is needed to clarify the merits of the differing indices available to justify aggregation and to clarify the frequency with which the indices yield differing conclusions.

Analysis of Multilevel Data

In analyzing multilevel data, researchers typically use one of three procedures: multilevel random-coefficient modeling (for example, HLM), WABA, or cross-level operator analyses. In this section, we highlight five key differences among the procedures.

1. *Does the procedure include guidelines for determining whether the set of variables is best treated at the unit, individual, or interdependent level?*

WABA researchers use WABA criteria to determine whether their variables are best conceptualized and analyzed as the properties of homogeneous groups, independent individuals, or interdependent individuals within groups. WABA researchers then analyze their data accordingly. This approach is unique to WABA. The two parts of WABA are inextricably linked:

Table 12.1. Summary of Procedures Used to Justify Aggregation of Lower-Level Data to Higher Units of Analysis.

Question	Two	Two	Two	Two	Three
1. Does the procedure assess variability within one unit at a time or across the entire sample?	Within	Across	Across	Across	Across
2. Is the procedure commonly used to decide between two patterns of variability (homogeneity or independence) or among three patterns (homogeneity, independence, or heterogeneity)?	Two	Two	Two	Two	Three
3. Does the procedure consider each variable separately or the entire set of variables?	Each separately	Each separately	Each separately	Each separately	Entire set
4. Where does the procedure draw the line in determining whether data should be aggregated?	.70 rule of thumb	Significant F-test	.70 rule of thumb	Significant F-test and E-test	Uncertain
5. Are the values obtained in using the procedure influenced by group size?	No	No	No	Yes	Uncertain

1. The determination that a data set should be analyzed between groups (wholes), within groups (parts), or between and within groups (equivocal)
2. The subsequent analysis of the data at the appropriate level of analysis (based on WABA criteria)

Thus, for WABA, the answer to the preceding question is yes: WABA does provide guidelines for determining at what level a set of variables should be treated.

HLM and CLOP analyses are quite different. HLM and CLOP do not provide guidelines for determining the level of a researcher's variables. HLM and CLOP, unlike WABA, are not designed to test the merits of aggregation; rather, they are designed to test cross-level models, in which one or more higher-level variables is posited to have a direct or moderating effect on lower-level variables (see Kozlowski & Klein, Chapter One, this volume). Before conducting HLM and/or CLOP, a researcher must determine whether a given variable should be aggregated to and analyzed at the unit level of analysis. Thus a researcher using HLM and/or CLOP typically relies on theory, r_{wg} , ICC(1), ICC(2), and/or eta-squared to justify the aggregation of some predictors to a higher unit level before analyzing the effects of these measures on a lower-level dependent variable.

2. Does the procedure allow a researcher to test multilevel models, assessing whether the relationship between two or more variables holds at two or more levels of analysis?

Multilevel models posit that a relationship between two variables holds at two or more levels of analysis (House, Rousseau, & Thomas-Hunt, 1995; Klein, Dansereau, & Hall, 1994; Rousseau, 1985; see also Kozlowski & Klein, Chapter One, this volume). Thus, for example, climate for service and performance exhibit a multilevel relationship if group climate for service is positively related to group performance and if organizational climate for service is positively related to organizational performance. It is possible to use HLM and CLOP in conjunction with r_{wg} , ICC(1), ICC(2), or eta-squared to test multilevel models, but WABA was specifically designed to test such models and thus does so explicitly and effectively.

To test multilevel relationships, WABA researchers conduct sequential analyses, first testing the relationship between two or more variables at one level of analysis in one equation and subsequently testing the relationship between the variables at a different level of analysis in a second equation. Consider again the relationship of climate to service and performance. To test whether the variables have the multilevel relationship that has been suggested, a WABA researcher would examine whether measures of climate for service and performance are homogeneous within groups and covary between groups (indicative of a group wholes relationship), and whether they are homogeneous within organizations and covary between organizations (indicative of an organizational wholes relationship). WABA allows a researcher to test more complex multilevel models as well. For example, WABA might be used to show that the relationship between two variables reflects not only between-group differences and within-group homogeneity (indicative of a group wholes relationship) but also group heterogeneity within organizations (indicative of a relationship reflecting group parts within organizations).

Multilevel models are complex conceptually, methodologically, and analytically. Yet another complexity is lexical. Most current organizational writers (e.g., House et al., 1995; Klein et al., 1994; Rousseau, 1985; see also Kozlowski & Klein, Chapter One, this volume) define a multilevel relationship as we have here: as a relationship between two or more variables that holds at more than one level of analysis. However, following Miller (1978), WABA researchers (e.g., Dansereau, Alutto, & Yammarino, 1984) refer to these relationships not as multilevel relationships but as *cross-level relationships*. As the next section indicates, however, most current organizational writers (e.g., House et al., 1995; Klein et al., 1994; see also Kozlowski & Klein, Chapter One, this volume) mean something quite different by that term. For clarity in the remainder of this chapter, we use the term *multilevel model* as we have defined it here, and we use the term *cross-level model* as we define it in the following sections.

3. Does the procedure allow a researcher to test cross-level direct-effect models, simultaneously assessing the effects of independent variables at different levels of analysis on a lower-level dependent variable?

A cross-level direct-effect model suggests that one or more higher-level variables has a main effect on a lower-level variable (House et al., 1995; Klein et al., 1994; Rousseau, 1985; see also Kozlowski & Klein, Chapter One, this volume). For example, the characteristics of an organization's reward system may influence individual employees' pay satisfaction. Both hierarchical linear modeling and cross-level operator analyses allow a researcher to test cross-level direct-effect models, simultaneously assessing the effects of independent variables at two or more levels of analysis (for example, organization and individual) on a lower-level (for example, individual) dependent variable. WABA, unlike cross-level operator and HLM analyses, does not allow a researcher to assess simultaneously the effects of independent variables at different levels of analysis. (Question 5, later in this section, highlights differences between HLM and CLOP analyses of cross-level direct-effect models.)

4. *Does the procedure allow a researcher to test cross-level moderator-effect models, assessing the extent and correlates of between-unit variability in the slope of the relationship between two lower-level variables?*

A cross-level moderator-effect model posits that the strength of the relationship between two lower-level (for example, individual-level) variables varies from unit to unit, typically as a function of particular unit characteristics. Thus, to take a simple example, the relationship between individual pay and individual performance (two individual-level variables) may be stronger in units that award pay raises for performance and weaker in units that award pay raises for seniority. A unit's pay policy is a unit-level variable that moderates the relationship between individual pay and individual performance.

HLM is explicitly designed to test cross-level moderator-effect models; that is, HLM is designed to identify whether lower-level slopes vary across higher-level units, and whether lower-level slope variability is related to higher-level variables. HLM is well suited to these types of analyses because it allows researchers to treat slope variation as a random factor. This is accomplished by including a specific error term for slope variation (see Hofmann, Griffin, & Gavin, Chapter Eleven, this volume). This "random effect" for slope variation is unique to HLM and is not estimated in either cross-level operator analyses or WABA. In HLM, cross-level inter-

action effects are assessed by determining the degree to which higher-level factors explain slope variation.

To some extent, CLOP and WABA can be used to estimate the consistency of slopes across groups. Using CLOP, one can test for slope differences by examining the block of interactions between group membership (coded as $J - 1$ dummy codes) and the independent variable. If the interaction is significant, it suggests that the slope between the independent and dependent variable varies across groups. Note that this test can be "expensive" in terms of degrees of freedom, particularly if J (the number of groups) is large because it requires an additional $J - 1$ degrees of freedom. In contrast, HLM performs this test using only two degrees of freedom—one for the estimate of slope variance, and one for the estimate of slope-intercept covariance. One may also use CLOP to test whether the slope of the relationship between a lower-level independent variable and the dependent variable varies as a function of a specific group characteristic (for example, group structure) by forming a moderator variable. This is not so "expensive" in terms of degrees of freedom, because the interaction term in this case has just one degree of freedom. In WABA, one can also test for slope variation among groups by either creating subsets or using a multiple regression WABA approach (see Dansereau & Yammarino, Chapter Ten, this volume). Nevertheless, it is worth reiterating that neither CLOP nor WABA explicitly allows for a random slope term; consequently, neither approach integrates tests of slope differences as explicitly or as efficiently as HLM does.

5. *In testing cross-level direct-effect models, does the procedure evaluate the effects of group-level predictors at the group level of analysis, or does the procedure evaluate the effects of group-level predictors at the individual level of analysis?*

This question captures key differences between the ways in which CLOP and HLM test both cross-level direct-effect models and cross-level moderating-effect models. (Because WABA does not allow researchers to test cross-level direct-effect models and was not designed to test cross-level moderating-effect models, this section focuses exclusively on CLOP and HLM.)

In testing cross-level models, HLM first estimates the relationships between the dependent variable and the lower-level (for

example, individual-level) independent variables within each unit. HLM then examines the relationship between the outcomes of this analysis (the unit intercepts and slopes) and the higher-level (that is, unit-level) independent variables. Thus HLM uses unit-level independent variables to explain between-unit variance in the dependent variable intercepts and slopes.¹ In contrast, CLOP uses both individual-level independent variables and unit-level independent variables to explain total (within- and between-unit) variance in the dependent variable. CLOP analyses do not partition variance into within- and between-unit components. This difference between CLOP and HLM has potentially important implications for estimated main-effect sizes, estimates of standard error, degrees of freedom, and interpretation of the results. We briefly consider each of these implications.

In HLM, higher-level independent variables—that is, independent variables varying solely between units—are used, as we have noted, to explain between-unit variability. Thus, in using HLM to test a cross-level direct-effect model, one may find that an organizational characteristic (for example, organizational climate) explains 60 percent of the between-organization variance in employee turnover intentions. This means that 60 percent of the variance among the organizational intercepts for employee turnover intentions is explained by the variable organizational climate. Organizational climate does not explain 60 percent of the total (within- and between-organization) variance in employee turnover intentions. If, for example, just 10 percent of the variance in employee turnover intentions lies between organizations, then organizational climate will explain approximately 6 percent of the total variance (60 percent of 10 percent of the total variance = 6 percent of the total variance).

In cross-level operator analyses, by contrast, higher-level independent variables are used to explain total variance (that is, between- and within-unit variance) in the dependent variable. Thus a researcher using cross-level operator analyses to assess the same cross-level direct-effect relationship between organizational climate and employee turnover intentions (within the same data set already described) will obtain results that look quite different from the results obtained by the researcher using HLM. Because CLOP analyses reveal the percentage of total variance explained (not the percentage of between-unit variance explained), CLOP reports effect sizes much smaller than in HLM. Thus, in the example just de-

scribed, organizational climate will explain approximately 6 percent of the total variance. Accordingly, great care must be taken in interpreting and comparing the results of CLOP and HLM. The finding that a variable explains 60 percent of the between-unit variance in a dependent variable may appear, at first glance, far more consequential than the finding that the variable explains 6 percent of the total variance in the dependent variable (Bryk & Raudenbush, 1992). In fact, both findings capture the same relationship between (in this example) climate and turnover intentions—the same relationship, assessed with different metrics.

The results obtained with HLM and with cross-level operator analysis differ not only in effect size but also in the standard error used in testing the significance of the cross-level effect. In cross-level operator analyses, standard error estimates (provided in the statistical output) are based on total variance because the analysis does not partition variance into the “between” and “within” components for computation of the standard errors. In HLM, by contrast, standard errors are based on partitioned variance: standard errors for lower-level variables (called level 1 variables in HLM) are based on lower-level (for example, within-group) variance, and standard errors for higher-level variables (called level 2 variables in HLM) are based on higher-level (for example, between-group) variance. One well-documented consequence is that results from cross-level operator analyses will, in general, provide standard error estimates for the higher-level variables that are too small (Bryk & Raudenbush, 1992; Kreft & de Leeuw, 1998). This means that cross-level operator analyses may show that a group-level variable was significantly related to the outcome variable, whereas HLM analyses of the same data and the same relationship may show that the group-level variable was not significantly related to the outcome. Standard error estimates for lower-level (level 1) variables are also affected, although the exact form of the bias is less well documented. In short, HLM and cross-level operator analyses calculate standard errors differently, and this can lead to quite different estimates of statistical significance.

CLOP and HLM also treat the degrees of freedom for higher-level variables differently. In HLM, the degrees of freedom for a higher-level variable are $J - 1$ (where J is the number of higher-level units). In cross-level operator analyses, the degrees of freedom for the higher-level variable are $N - 1$ (where N is the number of total

observations) when the higher-level variable is a single, metric variable (for example, a measure of a specific characteristic of higher-level units, such as group structure). When, however, the higher-level variable is a nominal measure of unit membership (dummy coded), then HLM and CLOP test the effects of the higher-level measure-of-unit measure, using comparable degrees of freedom (that is, $J - 1$).

In sum, a researcher who used both HLM and cross-level operator analyses to analyze the same cross-level relationship in the same data would be likely to find that the two analytical procedures yield very similar parameter estimates (regression coefficients) but different—perhaps quite different—effect sizes, standard errors, and significance levels. As we shall show in the third section of this chapter (and as discussed in Chapter Nine), it is not difficult to use the results from cross-level operator analyses to approximate the results that would have been obtained with HLM. Although this renders effect sizes from CLOP quite similar to those obtained with HLM, this does not obviate issues regarding statistical significance and standard error.

Summary Comments

Our comparison of WABA, CLOP, and HLM highlights important differences among the techniques (see Table 12.2). WABA was designed to identify the appropriate level or levels to describe the relationships among the variables in a data set. In contrast, CLOP and HLM were designed to test cross-level models. Unlike WABA, neither CLOP nor HLM incorporates any decision rules to determine the appropriate level or levels at which the relationships should be described. Researchers using CLOP and HLM must use other procedures, such as r_{wg} , ICC(1), ICC(2), and eta-squared, to justify the aggregation of their measures to the unit level. CLOP and HLM differ insofar as HLM allows one to model the non-independence or clustering present in the data; CLOP does not. Thus, in testing the effects of higher-level (for example, group) independent variables on a lower-level (for example, individual) dependent variable, HLM and cross-level operator analyses will yield very similar parameter estimates but different estimates of effect sizes, standard errors, and statistical significance. Finally, because HLM allows one to specify a random slope term, HLM is particularly effective in testing cross-level moderating-effect models. Understanding these key differences among the approaches can help one select an approach that is matched to one's theoretical question.

Table 12.2. Summary of Procedures Used to Analyze Multilevel Data.

Question	WABA	HLM	CLOP
1. Does the procedure include guidelines for determining whether variables are best treated at the unit, individual, or interdependent level?	Yes	No	No
2. Does the procedure allow a researcher to test multilevel models?	Yes, designed to do so	Yes, but not designed to do so	Yes, but not designed to do so
3. Does the procedure allow a researcher to test cross-level direct-effects models?	No	Yes	Yes
4. Does the procedure allow a researcher to test cross-level moderator-effect models?	Yes, but not designed to do so	Yes	Yes
5. In evaluating cross-level models, does the procedure evaluate the effects of group-level predictors at the group or individual level of analysis?	Not applicable	Group level of analysis	Individual level of analysis

Analyses of a Simulated Data Set

In this section, we report the results of our analyses of a simulated data set created by our coauthor Paul Bliese. The goal was to create a data set that would both be of interest to organizational researchers and illustrate similarities and differences among the statistical procedures commonly used to analyze multilevel data. Clearly, the complexity of the various multilevel approaches prohibits us from providing detailed descriptions of the results from each of the data-analytic techniques. Further, only a series of Monte Carlo studies could provide a definitive analysis of the similarities and differences among the multilevel procedures described in the preceding section of the chapter. Nevertheless, we believe that the analyses of the mock data set are helpful in clarifying similarities and differences among the multilevel data-analytic procedures. We begin by providing a detailed description of the data set. Next, we describe the results of several analyses of the merits of aggregating specific variables to the unit level. Finally, we present the results of WABA, HLM, and cross-level operator analyses of the simulated data set.²

Using a random-number generator, Bliese created a data set that contained 750 individual-level observations. The 750 observations were nested within fifty groups, and each group contained fifteen members. Thus the data were created to simulate a condition in which a researcher gathered measures of eight constructs (one dependent variable and seven predictors) from 750 employees nested within fifty groups of fifteen members each. Using variations of procedures discussed in Bliese and Halverson (1998) and Bliese (1998), Bliese created the data set, specifying target ICC(1) values and intercorrelations among the variables. Because Bliese used a random-number generator to create the data set, the actual ICC(1) values and intercorrelations among the variables in the data set do not conform precisely to his specifications. As with any simulated data set, one would expect the average ICC(1) values and intercorrelations to match the original specifications if the data set were repeatedly created. In the present case, however, the original specifications provided guidelines for how the data were created, and so the conformity of the data to the exact specifications is less important.

Variables

The simulated data set contains eight variables. The first variable was created to simulate an individual-level measure of Job Satisfaction, the dependent variable for the data set. This variable was created to vary both within and between groups, with an expected ICC(1) of .15 (actual ICC(1) = .11). The remaining variables were created as possible predictors of Job Satisfaction.

The first independent variable was created to simulate a measure of Cohesion. This variable was created to have an ICC(1) of .20 (actual ICC(1) = .22). The variable was created to have a group-level correlation of .50 with Job Satisfaction (that is, a correlation of .50 between the Cohesion group means and the Job Satisfaction group means) and a within-group correlation of zero with Job Satisfaction. The second independent variable was created to simulate a measure of Positive Affect. This variable was created to be unrelated to group membership, with an expected ICC(1) value of zero (actual ICC(1) = .02) and a raw correlation of .25 with Job Satisfaction.

The third independent variable was created to simulate a measure of Pay. This variable was created to have more within-group variability than between-group variability, with an expected negative ICC(1) value (the actual ICC(1) value was $-.01$, suggesting a weak frog-pond variable). The variable was created to have a within-group (or frog-pond) correlation of .40 with Job Satisfaction. The fourth independent variable was created to simulate a measure of Negative Leader Behaviors. The variable was created to have an ICC(1) value of .10 (actual ICC(1) = .04). Further, the variable was created to have a group-level correlation with Job Satisfaction of $-.70$ and a within-group (frog-pond) correlation with Job Satisfaction of $-.30$.

The fifth and sixth independent variables were created to simulate measures of Workload and Task Significance, respectively. Further, these variables were generated to interact with each other in their relationships with Job Satisfaction. Workload items were generated to have an ICC(1) of 0 (actual ICC(1) = 0), and Task Significance items were created to have an ICC(1) of .10 (actual ICC(1) = .11). Workload was created to have a correlation of $-.40$ with Job Satisfaction when Task Significance was low. When Task Significance was high, however, Workload was created to be unrelated to Job Satisfaction.

The seventh variable was created to simulate a measure of perceptions of the Physical Work Environment. Presumably there would be low variability among group members on a rating of this nature; consequently, this variable was generated to have an ICC(1) of .70 (actual ICC(1) = .67). The Physical Environment variable was created to have a group-level correlation of .40 with Job Satisfaction and a within-group correlation of 0 with Job Satisfaction.

The development of the mock data set was guided by several implicit hypotheses:

1. Job Satisfaction is positively related to Cohesion, Positive Affect, Pay, and Physical Work Environment
2. Job Satisfaction is negatively related to Negative Leader Behaviors and Workload
3. The relationship between Job Satisfaction and Workload is moderated by Task Significance.

Preliminary Analyses: Making the Decision to Aggregate

Table 12.3 lists each variable's mean, standard deviation, ICC (1), ICC(2), eta-squared, within-eta correlation, between-eta correlation, F-test, E-ratio, and inverse-F. Note that the eta-squared is simply the square of the between-eta correlation. Further, the F-test provides a test of the significance of eta-squared, the between-eta correlation, and ICC(1). The inverse-F is used by WABA researchers when a variable's within-eta correlation is larger than its between-eta correlation (see Dansereau et al., 1984; see also Dansereau & Yammarino, Chapter Ten, this volume, for further details).

Unfortunately, we cannot provide r_{wg} values for the variables in this data set. r_{wg} or more appropriately, $r_{wg,j}$, assesses variability among group members' responses to the survey items within a scale, but the data set simulates scale means, with no information whatsoever regarding mock scale items. That is, r_{wg} cannot be calculated in the absence of information regarding the number of items within a scale and the variability of item scores within each group.

Table 12.4 presents the raw-score, group-level (between-group), and within-group correlations between the independent variables and the dependent variable, Job Satisfaction. The raw-score correlation is the individual-level correlation between the 750 dependent and independent variables. The group-level (between-group)

Table 12.3. Means, Standard Deviations, and Summary Statistics for the Simulated Data Set.

Variable	M	SD	ICC(1)	ICC(2)	Eta-squared	Eta-between	Eta-within	F	E-ratio	Inverse F
Job satisfaction	0.08	2.48	.11	.65	.17	.41	.91	2.87**	.45	.35
Cohesion	0.23	2.51	.22	.81	.27	.52	.85	5.26**	.61	.19
Positive affect	0.06	0.98	.02	.28	.09	.30	.95	1.38**	.32	.72
Pay	0.04	1.01	-.01	-.27	.05	.23	.97	.79	.24	1.27
Negative leadership	0.00	1.09	.04	.38	.10	.32	.95	1.60**	.34	.62
Workload	0.00	1.04	-.00	-.07	.06	.25	.97	.93	.26	1.08
Task significance	-.01	1.01	.11	.65	.17	.41	.91	2.89**	.45	.35
Physical environment	0.07	1.43	.67	.97	.71	.84	.55	33.79***	1.52	

* $p < .05$; ** $p < .01$; *** $p < .001$

correlation is the correlation among the fifty group means, and the within-group correlation is the correlation of X and Y deviation scores. (See Dansereau & Yammarino, Chapter Ten, this volume, for the calculation of these correlations.) Intercorrelations among the independent variables are not presented because, given the way the data were generated, the correlations among the independent variables were near zero (the average intercorrelation among the independent variables was .002).

The preliminary analyses presented in Tables 12.3 and 12.4 provide the information necessary to compare different standards commonly used to justify aggregation to the unit level of analysis. As already noted, we cannot calculate r_{wg} values for the simulated data. However, we can compare the conclusions a researcher would make in basing the decision to aggregate on ICC(1), eta-squared, ICC(2), and WABA. The identical F-test is used to test the statistical significance of ICC(1) and of eta-squared. Thus, if a researcher relied solely on the presence of a statistically significant F (that is, on significantly greater between-group than within-group variance in the data) to determine whether to aggregate a measure to the

Table 12.4. Correlations Between Independent Variables and Satisfaction.^a

Variable	Raw Correlation	Between-Group Correlation	Within-Group Correlation
1. Cohesion	0.05	0.38**	-0.04
2. Positive affect	0.18**	0.21	0.17**
3. Pay	0.37**	-0.08	0.42**
4. Negative leadership	-0.29**	-0.61**	-0.25**
5. Workload	-0.25**	-0.32*	-0.24**
6. Task significance	0.02	-0.00	0.03
7. Physical environment	0.10*	0.33**	-0.03

^an = 750 for the individual-level correlations; 700 (750 - 50) for the within-group correlations; and 50 for the between-group correlations

*p < .05, two-tailed

**p < .01, two-tailed

group level of analysis, the researcher would aggregate six of the eight simulated variables to the group level: Job Satisfaction, Cohesion, Positive Affect, Negative Leader Behavior, Task Significance, and Physical Work Environment. The F-test for eta-squared, and for ICC(1), is statistically significant for each of these variables. If, instead, a researcher chose to aggregate only those measures showing reliable group differences, as evidenced by ICC(2) values over the conventional reliability standard of .70, then the researcher would aggregate just two of the eight variables: Cohesion and Physical Environment.

WABA combines the information presented in Table 12.3 (specifically, the between- and within-eta correlations) with the information presented in Table 12.4 (the between-group, within-group, and total correlations), to yield the information presented in Table 12.5. (See Dansereau & Yammarino, Chapter Ten, this volume, for more information regarding the calculations on which Table 12.5 is based.) The WABA researchers examined the information in these tables to determine at what level the data should be analyzed: at the level of wholes (that is, between groups), at the level of the equivocal condition (that is, between and within groups), or at the level of parts (that is, within groups).

Table 12.5. WABA-Components Analysis.

Variables and Relationships	Between Components	Within Components	Total Correlation	Inference from WABA
Satisfaction and				
Cohesion	.08	-.03	.05	Individual
Positive affect	.03	.15	.18	Individual
Pay	.01	.37	.37	Parts
Negative leadership	-.08	-.21	-.29	Individual
Workload	-.03	-.22	-.25	Individual
Task significance	-.00	.03	.02	Individual
Physical environment	.11	-.01	.09	Wholes

More specifically, the WABA researchers began by examining the F-ratio and E-test values (see Table 12.3) to compare the between-eta and within-eta correlations. The WABA researchers concluded that, with the exception of Physical Environment, the variables as a whole showed no clear indication of predominant between-group variation or of predominant within-group variation. Examining the relationships among the variables (shown in Table 12.4), as well as Z-tests and A-tests of these relationships (not shown in Table 12.4 but described in Chapter Ten), the WABA researchers made the following observations:

1. The relationship between Job Satisfaction and Pay is primarily a function of within-group differences.
2. The relationships between Job Satisfaction and Cohesion, Negative Leadership, and Physical Environment are primarily a function of between-group differences.

Given that Job Satisfaction is related to some variables primarily as a function of between-group differences and related to other variables primarily as a function of within-group differences, the WABA researchers concluded that the relationships among the variables are most appropriately conceptualized and analyzed at the individual level of analysis. WABA researchers define the pure individual level—the equivocal condition—as one that varies freely within and between groups.

Finally, the WABA researchers examined the “between” and “within” components underlying the raw correlations among the variables (see Table 12.5). The “between” components are formed by multiplying the between-eta correlations for x and y by the between-group correlation. The “within” components are formed by multiplying the within-eta correlations for x and y by the within-group correlation. The two components sum to the total individual-level, raw-score correlation. The findings in Table 12.5 reveal that only the within-group component for the correlation of Pay Satisfaction and Job Satisfaction supports a parts inference. Only the between-group component for the correlation of Physical Environment and Job Satisfaction supports a wholes inference. The relationships of the remaining variables and Job Satisfaction support an equivocal, or individual-level, inference. For WABA researchers, these results lend further support to the assertion that the variables in the data set vary both within and between groups and should thus be

conceptualized and analyzed at the individual level of analysis. Accordingly, the WABA analysts concluded that none of the variables in the mock data set should be aggregated to the unit level.³

Summary Comments

Researchers using different criteria to justify aggregation may indeed reach very different conclusions regarding the merits of aggregation. This could lead to differences in the specification of the theoretical model to be tested. Researchers relying on a statistically significant F-test for ICC(1) or eta-squared would report that six of the eight variables in the simulated data set could justifiably be aggregated. Researchers relying on ICC(2) values exceeding .70 would report that two of the eight variables could justifiably be aggregated. Researchers using WABA, however, determined that none of the variables should be aggregated.

Using WABA to Analyze the Relationship Between Predictors and Job Satisfaction

Building on the results just discussed, the WABA researchers analyzed the relationship between the independent variables and Job Satisfaction at the individual level of analysis. The WABA researchers did not aggregate any of the independent variables. This is an important detail because several of the measures—Cohesion, Negative Leader Behavior, Physical Environment, and Task Significance—were aggregated to the group level in the HLM and CLOP analyses. Thus the WABA analysts simply regressed Job Satisfaction on the raw, individual-level measures of all the independent variables. The results appear in the first column of Table 12.6.

The results indicate that three of the predictors—Positive Affect, Pay, and Physical Environment—are significantly positively related to Job Satisfaction. Two of the predictors—Negative Leader Behavior and Workload—are significantly negatively related to Job Satisfaction. Neither Cohesion nor Task Significance is significantly related to Job Satisfaction. Finally, the interaction of Task Significance and Workload is significant: the greater the Task Significance, the weaker the relationship between Workload and Job Satisfaction. Together, the eight predictors explain 29 percent of the total, individual-level variance in Job Satisfaction (that is, 29 percent of the total between- and within-group variance) in Job Satisfaction.

Table 12.6. Comparison of Unstandardized Parameter Estimates for the Equations Predicting Job Satisfaction.

Variable	WABA	HLM	CLOP
Positive affect (raw)	0.34**	0.31**	0.33**
Pay (raw)	0.82**	0.83**	0.83**
Workload (raw)	-0.42**	-0.34**	-0.35
Group cohesion (mean)	—	0.15*	0.20*
Cohesion (raw)	0.04	—	—
Group negative leadership (mean)	—	-1.17**	-1.06**
Negative leadership (raw)	-0.58**	-0.43**	-0.43**
Group physical environment (mean)	—	0.17*	0.15*
Physical environment (raw)	0.18**	—	—
Group task significance (mean)	—	-0.24	-0.34
Task significance (raw)	0.01	—	—
Group task significance by workload interaction (cross-level)	—	1.24**	1.28**
Task significance by workload interaction (raw)	0.22**	—	—

* $p < .05$

** $p < .01$

*** $p < .001$

Using HLM to Analyze the Relationship Between Predictors and Job Satisfaction

The HLM analyses of the relationship between the predictors and Job Satisfaction were guided by nine hypotheses. The HLM hypotheses are identical in direction to the hypotheses outlined previously. However, the HLM hypotheses differ from the previous ones insofar as the HLM hypotheses specify the level (individual or group) of each predicted relationship. Together, the hypotheses describe a cross-level model that specifies lower-level effects (hypotheses 1 through 4), cross-level main effects (hypotheses 5 through 8), and a cross-level moderating effect (hypothesis 9).

Hypotheses 1 to 4 focus on individual-level (or level 1) relationships:

H1. *Individual Positive Affect will be positively related to individual Job Satisfaction.*

H2. *Individual Negative Leader Behavior will be negatively related to individual Job Satisfaction.*

H3. *Individual Pay will be positively related to individual Job Satisfaction within groups so that the higher an individual's Pay, relative to the pay of other group members, the higher the individual's Job Satisfaction.*

H4. *Individual Workload will be negatively related to individual Job Satisfaction.*

Hypotheses 5 to 8 focus on cross-level main effects. Analyses testing these hypotheses examine intercepts as outcomes (see Hofmann et al., Chapter Eleven, this volume).

H5. *After controlling for the individual-level predictors, Group Cohesion (that is, the group mean for Cohesion) will be positively related to the average Job Satisfaction of group members.*

H6. *After controlling for the individual-level predictors (including Negative Leader Behaviors), Group Negative Leader Behaviors (that is, the group mean for Negative Leader Behaviors) will be negatively related to average level of Job Satisfaction of group members above and beyond the individual effects of Negative Leader Behaviors on individual Job Satisfaction.*

H7. *After controlling for the individual-level predictors, Group Physical Work Environment (that is, the group mean of perceptions of the Physical Work Environment) will be positively related to average level of Job Satisfaction of group members.*

H8. *After controlling for the individual-level predictors, Group Task Significance (that is, the group mean of Task Significance) will be positively related to average level of Job Satisfaction of group members.*

Finally, hypothesis 9 tests whether a group-level predictor moderates the relationship between two individual-level variables. The analysis tests a cross-level moderating effect, examining slopes as outcomes (again, see Hofmann et al., Chapter Eleven, this volume):

H9. *Group Task Significance (that is, the group mean of Task Significance) will moderate the relationship between individual Workload and individual Job Satisfaction. Specifically, whereas individual Workload is expected to be negatively related to individual Job Satisfaction, this relationship will be weak in groups in which Task Significance is high.*

In the WABA analyses, as we have noted, none of the measures was averaged and aggregated to the unit level. Guided first by the implicit theoretical model that determined the creation of the mock data set, and then by supporting ICC(1) and ICC(2) values for the measures, the HLM researchers aggregated (averaged) four measures to the unit level within the HLM analyses: Cohesion, Negative Leader Behaviors, Physical Work Environment, and Task Significance. In the remainder of the chapter, we refer to the aggregated measures as *Group Cohesion*, *Group Negative Leader Behaviors*, *Group Physical Work Environment*, and *Group Task Significance*, respectively. Further, within the HLM analyses, the effects of Negative Leadership Behavior were tested both at the individual level of analysis (Negative Leadership Behaviors) and at the group level of analysis (Group Negative Leader Behaviors).

A total of four models was specified to test the nine hypotheses just identified. The HLM analysts opted for this sequence because it provided a minimum number of models for testing the hypotheses and ultimately resulted in a final model that simultaneously accounted for all the hypothesized effects. The first model run by the HLM analysts was a null model with no predictors. This model partitioned the variance in Job Satisfaction into between- and within-group components. The second model was an individual-level model specifying the four individual-level predictors of Job Satisfaction but containing no predictors at the group level. The third model was a group-level model that included the four individual-level predictors from the previous model as well as the four group-level predictors of Job Satisfaction. The fourth and final model included the specified cross-level interaction in addition to the four previously identified individual-level predictors and the four previously identified group-level predictors.

Null model. The null model specified no individual-level or group-level predictors, only Job Satisfaction as an outcome. This model

revealed that 11 percent of the variance in Job Satisfaction resided between groups; note that this is equivalent to the ICC(1) reported earlier. Although this certainly indicates that the majority of the variance in Job Satisfaction lies at the individual level, the group-level variance component is significant ($\chi^2 = 140.78$, 49 df, $p < .001$) and suggests that exploration of its antecedents is a worthwhile pursuit.

Individual-level model. Next, the HLM analysts modeled the four individual-level predictors of Job Satisfaction. These predictors were Positive Affect, Negative Leader Behaviors, Pay, and Workload. In specifying this model, all predictors with the exception of Pay were centered around their grand means. Pay was centered around its group mean because it was hypothesized to operate as a within-group, or frog-pond, effect in which the effect of an individual's Pay on his or her Job Satisfaction was dependent on his or her level of Pay relative to the pay of other group members. These centering decisions are consistent with earlier discussions (see Hofmann et al., Chapter Eleven, this volume) and with the recommendations of Hofmann and Gavin (1998).

In this model, each of the coefficients is significant in the anticipated direction: Positive Affect ($\beta_{1j} = .30$, $t\text{-ratio} = 4.01$, $p < .001$), Negative Leader Behaviors ($\beta_{2j} = -.49$, $t\text{-ratio} = -6.19$, $p < .001$), Pay ($\beta_{3j} = .85$, $t\text{-ratio} = 10.31$, $p < .001$), and Workload ($\beta_{4j} = -.39$, $t\text{-ratio} = -3.39$, $p < .01$). Collectively, the four predictors account for 38 percent of the within-group variance in Job Satisfaction. For two reasons, this estimate of variance explained differs from the estimate of variance explained in the regression performed by the WABA researchers:

1. The WABA researchers included all the predictor variables in their regression, whereas the HLM researchers included only the four variables (already noted) in this model.
2. The WABA researchers assessed the percentage of total (within- and between-group) variance explained, whereas the HLM analysis assessed the percentage of within-group variance explained.

One other finding of interest from the HLM model is that the slope of the relationship between Workload and Job Satisfaction varies significantly across groups ($\chi^2 = 100.94$, 49 df, $p < .001$). This between-group variability in the slope of the Workload-Job Satisfaction

relationship is the focus of hypothesis 9. None of the other predictors had slope effects that varied across groups.

Group-level model. Next, the four group-level predictors of Job Satisfaction (Group Cohesion, Group Negative Leader Behaviors, Group Physical Work Environment, and Group Task Significance) were added to the previous model. The level 1 (here, the individual level) portion of the model remained the same in terms of the predictors included and the centering options specified. Each of the group-level predictors in this cross-level main-effect model was significant: Group Cohesion ($\gamma_{01} = .16$, $t\text{-ratio} = 2.23$, $p < .05$), Group Negative Leader Behaviors ($\gamma_{02} = -1.17$, $t\text{-ratio} = -4.23$, $p < .001$), Group Physical Work Environment ($\gamma_{03} = .18$, $t\text{-ratio} = 2.42$, $p < .05$), and Group Task Significance ($\gamma_{04} = -.51$, $t\text{-ratio} = -2.31$, $p < .05$). As a set, the four group-level predictors account for 45 percent of the between-group variance in Job Satisfaction. There is still a significant amount of remaining variance between groups in Job Satisfaction ($\chi^2 = 78.41$, 45 df, $p < .01$). However, there are no other group-level predictors in the data set to model this remaining variance.

Full model. The final model builds on the previous model by adding into the model a test of the cross-level moderating effect of Group Task Significance on the Workload-Job Satisfaction relationship. The results of this model are shown in the second column of Table 12.6. (The table also includes the final coefficient estimates for all the variables in the model.) Of specific interest is the significant cross-level moderating effect; see "Group Task Significance by Workload Interaction (Cross-Level)" in Table 12.6. This finding indicates that when Group Task Significance is high, the negative within-group relationship between Workload and Task Significance is attenuated. The pattern of significance for the other coefficients is the same as in the preceding models except that in this model the main effect of Group Task Significance is no longer significant. Although a significant 68 percent of the variance in the Workload-Job Satisfaction relationship was successfully modeled with group Task Significance, significant variance remains across groups in that relationship ($\chi^2 = 65.21$, 48 df, $p < .05$), variance that could be modeled with other group-level predictors, theory and data permitting.

Using CLOP to Analyze the Relationship Between Predictors and Job Satisfaction

As we have noted, HLM and cross-level operator analyses are quite similar. Both are designed to test cross-level models. They differ, however, in their partitioning of the variance in the lower-level dependent variable (and thus in their modeling of non-independence or clustering in the data). HLM evaluates the effects of lower-level independent variables on lower-level outcomes and the effects of higher-level independent variables on higher-level variance in intercepts and slopes. Traditional cross-level operator analyses, in contrast, predict total variance in the dependent variable. Thus, as we have noted, cross-level operator analyses assess the extent to which both individual- and group-level predictors explain total (within- and between-group) variability in the dependent variable. The analyses in the first part of this section illustrate traditional cross-level operator analyses.

As explained in Chapter Nine (James & Williams, this volume), analysts may take the results of traditional cross-level operator analyses and estimate the effects of higher-level independent variables on higher-level variance in intercepts and slopes, thereby approximating the results of hierarchical linear modeling. The analyses in the second part of this section illustrate this strategy. To facilitate the comparison of cross-level operator analyses and HLM, the CLOP researchers patterned their analyses on the HLM analyses just reported.

Traditional CLOP analyses. As the first step in the traditional cross-level operator analyses, the researchers regressed Job Satisfaction on the four individual-level predictors: Positive Affect, Negative Leader Behaviors, Pay, and Workload. (Following the HLM researchers, the CLOP researchers centered all predictors around their grand means with the exception of Pay, which was group-mean-centered.) Each of the coefficients is significant in the anticipated direction: Positive Affect ($\beta = .35$, $t\text{-ratio} = 4.38$, $p < .001$), Negative Leader Behaviors ($\beta = -.59$, $t\text{-ratio} = -8.27$, $p < .001$), Pay ($\beta = .87$, $t\text{-ratio} = 10.88$, $p < .001$), and workload ($\beta = -.43$, $t\text{-ratio} = -5.72$, $p < .001$). The beta coefficients are similar (though not identical) to those from the HLM analysis. Collectively, the four predictors account for 27.6 percent of the total (within- and between-group) variance in Job Satisfaction.

As the next step, the researchers assigned the group-mean scores for Cohesion, Negative Leadership, Task Significance, and Physical Environment back to the individual and used these scores as predictors. The four cross-level predictors are significantly related to Job Satisfaction: Group Cohesion ($\beta = .23$, $t\text{-ratio} = 3.90$, $p < .001$), Group Negative Leadership ($\beta = -1.07$, $t\text{-ratio} = -4.58$, $p < .001$), Group Work Environment ($\beta = .17$, $t\text{-ratio} = 2.73$, $p < .01$), and Group Task Significance ($\beta = -.39$, $t\text{-ratio} = -2.13$, $p < .05$). Together, the four cross-level predictors explain an additional 5.2 percent of the variance in Job Satisfaction, over and above the variance explained by the individual-level predictors. Thus the group- or cross-level predictors and the individual-level predictors explain a total of 32.8 percent of the variance in Job Satisfaction.

As the final step of the traditional cross-level operator analyses, the researchers examined the hypothesized cross-level moderating effect of Group Task Significance on the Workload-Job Satisfaction relationship. The interaction term is statistically significant ($\beta = 1.27$, $t\text{-ratio} = 7.45$, $p < .001$), explaining an additional 4.7 percent of the total variance in Job Satisfaction. Thus, all together, the individual-level predictors, the group-level predictors, and the interaction of Group Task Significance and Workload explain 37.5 percent of the total variance in Job Satisfaction. The parameter estimates for the full regression model are shown in Table 12.6. Note that the total variance explained by the CLOP regression (37.5 percent) exceeds the total variance explained by the WABA regression (29 percent). This reflects the fact the CLOP researchers aggregated four of the variables to the group level, whereas the WABA researchers did not. These four variables (Cohesion, Negative Leadership, Task Significance, and Physical Environment) were created in the mock data set to covary with Job Satisfaction more strongly between than within groups. Therefore, the aggregated variables had more predictive power (in the CLOP analysis) than the raw, disaggregated measures had (in the WABA analysis).

"Translating" the results of traditional CLOP analyses to estimate between-group (level 2) variance. The CLOP researchers performed a series of additional analyses to estimate the extent to which the group-level variables explained between-group variability (not total vari-

ability) in Job Satisfaction. Their first step was simply to obtain an estimate of the extent of between-group (versus within-group) variability in Job Satisfaction. To do this, the CLOP researchers regressed Job Satisfaction on $J - 1$ dummy codes representing Group Membership (J equals the number of groups). This model is identical to a one-way ANOVA (Cohen & Cohen, 1983). It partitions Job Satisfaction variance into its between-group and within-group components. The R-Squared value for this model is .17, suggesting that about 17 percent of the variation in Job Satisfaction resides between groups. It is important to recognize that this estimate is a slightly biased overestimate (as is its ANOVA analogue, eta-squared). It is an overestimate because the regression model includes forty-nine independent variables (the dummy codes for the fifty groups) but does not correct for this large number of predictors. A better estimate of the variance "explained by" Group Membership is provided by the shrunken R-square (Cohen & Cohen, 1983). This measure adjusts for the number of predictors. The shrunken R-square estimate is .11, or approximately 11 percent of the variance. Not coincidentally, the shrunken R-square estimate of .11 matches the ICC(1) estimate of .11 and the ICC estimate from the HLM analysis.

The next step for the CLOP analysts was to estimate the percentage of the between-group variance accounted for by the group-level predictors (Group Cohesion, Group Negative Leadership, Group Task Significance, and Group Physical Environment). Recall that the group-level predictors explained an additional 5.2 percent of the variance in Job Satisfaction, over and above the variance explained by the individual-level predictors. This 5.2 percent of the total variance represents 47 percent of the between-group variance ($5.2/11 = 47$). Note that this "translation" of the results of traditional CLOP analyses does not change the parameter estimates for the predictor variables, but only the estimate of the percentage of variance explained. In comparison, the HLM analysis showed that the group-level predictors explain 45 percent of the between-group variance in Job Satisfaction.

Finally, to estimate the extent to which Task Significance explained between-group variability in the slope of the Workload-Job Satisfaction relationship, the researchers first needed an estimate of between-group variability in this relationship. To estimate between-group variability in the Workload-Job Satisfaction relationship, the

researchers conducted a three-step hierarchical regression. In the first step, they regressed Job Satisfaction on Workload. In the second step, they added the forty-nine ($J - 1$) dummy codes for Group Membership to the model. In the third step, they added the forty-nine Workload by Group Membership interaction terms to the model. At each step, the significance of the block was assessed. The important step in the detection of cross-level interaction effects is the third block. This block accounts for significant variance [$F(49,650) = 2.84, p < .001$], indicating that the relationship between Workload and Job Satisfaction varies significantly between groups. The R-square for the model with only Workload is .06. The R-square for the model with Workload and Group Membership is .22, with a shrunken R-square of .16. Finally, the R-square for the third model, including the interaction term and both main effects, is .36, with a shrunken R-square of .26. Thus the researchers concluded (using the shrunken R-square) that the interaction of Group Membership and Workload accounts for 10 percent (.26 minus .16) of the total individual-level variance in Job Satisfaction; that is, between-group variability among the slopes explains 10 percent of the total variance in Job Satisfaction.

The traditional CLOP analyses, reported in the preceding subsection, showed that the interaction of Group Task Significance and Workload explained 4.7 percent of the total variance in Job Satisfaction. Accordingly, the CLOP researchers concluded that Group Task Significance accounts for 47 percent ($4.7/10$) of the between-group variation in the Workload-Job Satisfaction relationship. (As in the previous "translation," this translation of the results does not change the parameter estimates for the predictor variables.) In contrast, the results of the HLM analyses suggest that Group Task Significance explained 68 percent of the variation in this relationship.

Summary of Analyses

The results of each of the three sets of analyses, summarized in Table 12.6, provide strong support for the hypotheses. Of course, the data were developed to illustrate the hypotheses, so this is hardly a surprising result! More intriguing is a comparison of the results of the three analytical techniques. A critical consideration,

of course, is that the WABA results are based on the raw, individual-level data; the WABA researchers aggregated no variable to the group level. In contrast, both the HLM researchers and the CLOP researchers aggregated four variables to the group level: Cohesion, Negative Leadership Behaviors, Physical Environment, and Task Significance. Before discussing in more detail the differing results in Table 12.6, we must caution again that the results are illustrative, not definitive. They are based on one mock data set. Our three analyses of this data set are a useful beginning, not a final statement.

We begin our discussion with a comparison of the HLM and CLOP results. As the results in Table 12.6 indicate, the parameter estimates for the two models are very similar. This is to be expected because estimates of these "fixed" effects tend to vary little across estimation technique (Bryk & Raudenbush, 1992). The parameter estimates do differ somewhat in statistical significance, particularly for the group-level variables. For example, in the CLOP analysis, Group Cohesion is significant with a p value less than .01 (the exact p -value is .0005), but the p value in the HLM analyses is significant, with a p value of .036. This difference reflects the fact that HLM adjusts the standard errors for group-based non-independence (see Bliese, Chapter Eight, this volume, for a discussion of this form of non-independence), whereas CLOP makes no such adjustment. As we and many others have noted, this difference between HLM and CLOP may cause standard errors for the group-level variables in CLOP analyses to be too small, thus leading to Type I errors.

The HLM analyses and the traditional CLOP analyses do yield very different estimates of the variance explained by the predictors. This reflects the fact that HLM uses higher-level (for example, group) predictors to explain higher-level (for example, between-group) variance in intercepts and slopes, whereas traditional cross-level operator analyses use higher-level predictors to explain total variance in the dependent variable. While one may "translate" the results of traditional cross-level operator analyses to obtain estimates of the between-group variance explained, the results are similar but not identical to those obtained in using HLM. Further, it is not entirely clear how to evaluate the statistical significance of the "translated" (and thus transformed) CLOP results.

Researchers must use great care in interpreting the differing estimates of explained variance that HLM and CLOP provide. For

example, the HLM analyses indicate that the four group-level predictors explain 45 percent of the between-group variance in the Job Satisfaction. The traditional cross-level operator analyses, in contrast, indicate that the same four group-level predictors explain just 5.2 percent of the total variance in Job Satisfaction. The difference lies neither in the data nor in the relationships among the variables but in the reference point. The HLM analyses are conducted "with reference to" (that is, for the dependent variable of) the Job Satisfaction group intercepts; the CLOP analyses are conducted "with reference to" (that is, for the dependent variable of) the raw individual Job Satisfaction scores. The correct reference point depends in part on the researcher's theory. HLM analyses are commonly viewed as statistically superior to CLOP analyses; HLM analyses assess cross-level effects, using statistically appropriate degrees of freedom and estimates of standard error. HLM analyses may be difficult to interpret, however, if the between-unit intercepts of the dependent variable are not conceptually meaningful. (For further discussion of the meaning of aggregated variables, such as the group intercepts in HLM, see Kozlowski & Klein, Chapters One, this volume; Bliese, Chapter Eight, this volume; and Dansereau & Yammarino, Chapter Ten, this volume.)

The WABA analyses stand in rather sharp contrast to the HLM and CLOP analyses. A key difference, of course, is that the WABA researchers chose not to aggregate any of the independent variables; rather, the WABA researchers concluded that the relationships among the set of variables were best conceptualized and analyzed at the individual level of analysis. The WABA parameter estimates for the individual-level predictors are quite similar to the HLM and CLOP parameter estimates for these measures. The parameter and effect-size estimates for the variables left disaggregated by the WABA researchers, but aggregated by the HLM and CLOP researchers, are of course quite different. WABA, as we have noted, was designed to identify the single level or levels at which a set of relationships is best conceptualized and analyzed. WABA may analyze a set of relationships at "single levels" insofar as WABA allows one to conduct such analyses sequentially, but not simultaneously. In this sense, WABA is perhaps more constraining than HLM or CLOP, but WABA does allow a researcher to avoid at least some of the conceptual and statistical complexities that can arise in inter-

preting the results of HLM and CLOP analyses. In the present case, for example, it is relatively simple, both conceptually and statistically, to interpret the end results of the WABA analyses as indicative of the influence of a set of individual-level perceptual variables on individual-level Job Satisfaction. In contrast, the HLM and CLOP analyses suggest that Job Satisfaction is the product of both group and individual characteristics and, further, that the relationship between Job Satisfaction and one of its individual-level predictors (Workload) varies from group to group, in part as a function of the group's Task Significance.

Concluding Comments

In this chapter, we have compared and demonstrated a set of procedures commonly used to justify aggregation (r_{wg} , WABA, ICC(1), ICC(2), eta-squared) and a set of procedures commonly used to analyze multilevel data (WABA, HLM, and CLOP). We hope that we have, in the process, clarified commonalities and differences among these procedures.

The procedures used to justify aggregation are most likely to lead to the same conclusion in relatively unambiguous cases—that is, when a measure's variability lies quite predominantly between units, or when a measure's variability lies quite predominantly within units, or when a measure shows random variability within and between units. In more ambiguous cases, the procedures used to justify aggregation may yield differing conclusions.

Ultimately, aggregation is a judgment call informed by theory and data. Theory provides—or, at least, should provide—a detailed and convincing explanation of how and why lower-level measures may be combined to yield meaningful measures of unit characteristics. Each of the procedures used to justify aggregation provides a slightly different piece of evidence that the theory is correct—that aggregation is warranted. For example, r_{wg} provides evidence of the extent to which each of the units in a sample is characterized by within-unit agreement on the measure in question. Eta-squared and ICC(1) provide evidence of the extent to which the sample as a whole varies within and between units on the measure in question. ICC(2) provides evidence of the stability or reliability of the unit means on the measure in question. WABA provides evidence of the

extent to which not one measure but all the measures in a sample vary and covary primarily within units, between units, or within and between units. The pieces of evidence are related but by no means identical.

Within-and-between analysis, hierarchical linear modeling, and cross-level operator analyses allow researchers to examine the relationships among higher-level and lower-level measures. When higher-level variables are derived from lower-level measures, aggregation decisions have profound implications for the conceptualization and analysis of multilevel and cross-level models. We drew a rather sharp distinction between the first and second parts of this chapter. The third part of this chapter—the analysis of the mock data set—shows how integrally related aggregation decisions may be to the design and testing of multilevel and cross-level models. The differing aggregation decisions made by the WABA researchers, by contrast with the decisions made by the CLOP and HLM researchers, led to sharp differences in the researchers' respective analyses and interpretations of the data.

Even in the absence of differing aggregation decisions, however, WABA, HLM, and CLOP are quite distinct. WABA, we have emphasized, was designed to identify the level or levels describing a set of relationships within a data set. WABA is thus particularly well suited to the testing of multilevel theoretical models, as we have used that term in this chapter. HLM and CLOP, in contrast, were designed to test cross-level models, as we have used that term in this chapter, and are particularly well suited to this task. HLM offers statistical advantages that may be important in testing cross-level moderating-effect models.

We hope that our examination of the commonalities and differences among multilevel analytical techniques will assist organizational researchers in using these techniques. There is no one best technique for analyzing multilevel data. The procedures, we have shown, provide differing pieces of evidence, which may be used to answer differing kinds of theoretical questions. This chapter has described the current state of the art in multilevel data analysis, not the end state. As organizational scholars devote increasing attention to multilevel issues, our theoretical models will grow in sophistication, complexity, and variety (see Kozlowski & Klein, Chapter One, this volume). And with these changes will come the need for, and surely the development of, new ways to test multilevel

theory. As authors, we do not look forward to when this chapter is out-of-date; as researchers, however, we do.

Notes

1. Note that individual-level independent variables may explain not just within-unit variance but also between-unit variance in the dependent variable if the individual-level independent variable itself varies both within and between units. If an individual-level independent variable is group-mean centered, it varies only within units and thus can only explain within-unit variance in the dependent variable. See Hofmann and colleagues, Chapter Eleven, this volume, for further discussion of these issues.
2. The simulated data set is available on the Web site of the Society for Industrial and Organizational Psychology: <http://www.sio.org>. Individuals may download the data set and analyze it with WABA, CLOP, and HLM to gain hands-on experience in multilevel data analysis.
3. The WABA researchers also examined the homogeneity of the data set by splitting the sample into two groups: those with more within-group agreement on the dependent variable, and those with less within-group agreement on the dependent variable. They then conducted additional WABA analyses on the two subsamples. In the first subsample (characterized by relatively high within-group variability), none of the between-group correlations differed from the within-group correlations. In the second subsample (characterized by relatively low within-group variability), three of the between-group correlations were significant and differed from the analogous correlations in the high-variability subsample. On the basis of these results, the WABA researchers concluded that the sample was not homogeneous and thus that aggregated analyses would not represent the data. The WABA researchers believed that in these circumstances the best that could be done was to use the individual scores, recognizing that this was a violation of the homogeneity requirements of most statistical analyses.

References

- Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. *Psychological Bulletin*, 83, 762–765.
- Bliese, P. D. (1998). Group size, ICC values, and group-level correlations: A simulation. *Organizational Research Methods*, 1, 355–373.
- Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349–381). San Francisco: Jossey-Bass.

- Bliese, P. D., & Halverson, R. R. (1998). Group size and measures of group-level properties: An examination of eta-squared and ICC values. *Journal of Management*, 24, 157-172.
- Brown, K. G., Kozlowski, S.W.J., & Hattrup, K. (1996, August). *Theory, issues, and recommendations in conceptualizing agreement as a construct in organizational research: The search for consensus regarding consensus*. Paper presented at annual meeting of the Academy of Management, Cincinnati, OH.
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.
- Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, 83, 234-246.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Erlbaum.
- Dansereau, F., Alutto, J. A., & Yammarino, F. J. (1984). *Theory testing in organizational behavior: The variant approach*. Englewood Cliff, NJ: Prentice-Hall.
- Dansereau, F., & Yammarino, F. J. (2000). Within and Between Analysis: The variant paradigm as an underlying approach to theory building and testing. In K. J. Klein & S.W.J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 425-466). San Francisco: Jossey-Bass.
- Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Theoretical and methodological implications for organizational science. *Journal of Management*, 23, 623-641.
- Hofmann, D. A., Griffin, M., & Gavin, M. (2000). The application of hierarchical linear modeling to organizational research. In K. J. Klein & S.W.J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 467-511). San Francisco: Jossey-Bass.
- House, R., Rousseau, D. M., & Thomas-Hunt, M. (1995). The meso paradigm: A framework for the integration of micro and macro organizational behavior. In L. L. Cummings & B. M. Staw (Eds.), *Research in organizational behavior* (Vol. 17, pp. 71-114). Greenwich, CT: JAI Press.
- James, L. R., Demaree, R. J., & Wolf, G. (1984). Estimating within group interrater reliability with and without response bias. *Journal of Applied Psychology*, 69, 85-98.
- James, L. R., & Williams, L. J. (2000). The cross-level operator in regression, ANCOVA, and contextual analysis. In K. J. Klein & S.W.J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 382-424). San Francisco: Jossey-Bass.
- 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.
- Kozlowski, S.W.J., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology*, 77, 161-167.
- Kozlowski, S.W.J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S.W.J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 3-90). San Francisco: Jossey-Bass.
- Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.
- Miller, J. G. (1978). *Living systems*. New York: McGraw-Hill.
- Rousseau, D. (1985). Issues of level in organizational research: Multilevel and cross-level perspectives. In L. L. Cummings & B. M. Staw (Eds.), *Research in organizational behavior* (Vol. 7, pp. 1-37). Greenwich, CT: JAI Press.
- Schmidt, F. L., & Hunter, J. E. (1989). Interrater reliability coefficients cannot be computed when only one stimulus is rated. *Journal of Applied Psychology*, 74, 368-370.