From Micro to Meso: Critical Steps in Conceptualizing and Conducting Multilevel Research

KATHERINE J. KLEIN
University of Maryland

STEVE W. J. KOZLOWSKI
Michigan State University

Although interest in multilevel organizational theory, research, and methods has been on the rise in recent years, vigorous debates in the literature regarding appropriate ways to conceptualize and measure multilevel constructs, justify aggregation, and analyze multilevel models have contributed to confusion. New investigators interested in testing multilevel theory are intrigued, but wary. The goal of this article is to cut through the confusion, identifying the critical choices and issues a researcher may confront as he or she shifts from a single level to a multilevel perspective. The authors address four primary choices—construct and measurement issues, model specification, research design and sampling, and data analyses—describing critical steps in conceptualizing and conducting multilevel research.

As multilevel theoretical models and analytical systems gain increasing prominence in the organizational literature, many micro-organizational scholars may be contemplating a shift from purely micro-individual-level research to meso- or multilevel research (House, Rousseau, & Thomas-Hunt, 1995). We regard such a shift as laudable, indeed, essential; multilevel research is—at its best—complex, rigorous, and able to capture much of the nested complexity of real organizational life. And yet, we recognize that a shift from single-level to multilevel research may be daunting. The literature, with its many debates regarding the appropriate way to conceptualize multilevel models, jus-

---

Authors' Note: Material for this essay has been informed by chapters appearing in K. J. Klein & S.W.J. Kozlowski (Eds.) (2000), *Multilevel theory, research and methods in organizations: Foundations, extensions, and new directions* published by Jossey-Bass. In particular, issues relating to construct conceptualization, measurement, and sampling are based on Kozlowski and Klein (2000), and issues relating to aggregation and multilevel analyses are based on Klein et al. (2000). More detailed discussions of the issues addressed in this article can be found in the book and its chapters.

We thank the three anonymous reviewers for their comments and suggestions. We are also grateful to Larry Williams, Paul Bliese, Fred Dansereau, Mark Gavin, Mark Griffin, David Hofmann, Larry James, and Fran Yammarino. Our conversations with all of these people have greatly enhanced our understanding of multilevel research and data analysis.

Organizational Research Methods, Vol. 3 No. 3, July 2000 211-236

tify aggregation, and evaluate multilevel models, may engender more confusion than clarity (Kozlowski & Klein, 2000). We hope to cut through the confusion, providing guidance with respect to theory, research methods, and analyses.

Our goal in this essay is to facilitate the transition from single-level to multiple-level research by clarifying the choices a researcher faces on the road from micro- to meso–theory testing. We do not attempt to cover every multilevel issue, model, type of construct, or analysis. Rather, we focus on fundamental multilevel theoretical concerns and the alignment of these theoretical concerns with construct measurement, research design, and data analysis. More specifically, we focus on four key choices:

1. Construct and measurement choices: What is the nature of each higher level construct and how should each construct be operationalized?
2. Model choices: What kind of model describes the predicted relationships among the constructs?
3. Sampling choices: What kind of sample is necessary to operationalize key constructs and to test the predicted relationships?
4. Analysis choices: What analytical procedures may be used to test the predicted relationships?

Illustrative Example

To illustrate the nature and importance of these choices, we have created a hypothetical researcher, Dr. Faust, who seeks to develop a multilevel conceptualization of his heretofore single-level, micro ideas. We will examine Dr. Faust's "research" throughout the essay, referring repeatedly to the challenges and choices that Faust faces in shifting from a single-level conceptualization of the antecedents of individual performance to a multilevel conceptualization of the antecedents of individual and team performance. We hope that our use of this example provides this article with a unifying and practical focus. Our illustration focuses on individuals as lower level units and teams as higher level units to simplify the presentation. The issues we address, however, are applicable to the conceptualization and study of any human system composed of hierarchically nested levels.

A Few Specifics

Before addressing these issues, we provide additional background about Dr. Faust's research challenge. Faust has long been interested in the extent to which an individual's performance is predicted by his or her (a) self-efficacy, (b) non–work responsibilities (e.g., childcare, elder care), and (c) perceptions that his or her supervisor is a charismatic leader. Faust recently asked the president of the Acme Company to allow him to distribute surveys measuring these variables to Acme employees. In return, Faust promised to provide Acme with a feedback report summarizing his findings. Acme's president agreed to allow Faust to distribute the surveys to Acme's 1,000 employees but told Faust that Acme has, in recent years, instituted approximately 100 proximal (face-to-face) and distributed (virtual) teams. Acme's president thus hopes that Faust's research will shed light on the influence of team characteristics on both individual and team performance.

As Faust shifts from a single-level conceptualization of the antecedents of individual performance to cross-level and multilevel conceptualizations of the antecedents of

individual and team performance, he will have to alter numerous facets of his research, including the underlying theoretical model guiding the research. Faust's journey thus illustrates key phases of multilevel theory-building and theory-testing. We begin our examination with the most basic question of all: Why is this journey necessary?

The Relationship of Individual-Level Findings to Higher Level Findings

To illuminate the determinants of team performance, should Faust simply conduct his individual-level research as planned, analyze the results, and generalize them to the team level of analysis? The answer is no. To blindly generalize findings across levels of analysis is to commit a fallacy. Findings at one level of analysis do not generalize neatly and exactly to other levels of analysis, except under very restrictive circumstances (Firebaugh, 1979). When macro researchers attempt to generalize findings from aggregated data back to the lower level at which it was collected, they commit the well-known ecological fallacy. For example, school researchers may discover that aggregate student ability is highly related to student achievement at the school level. This finding may not, however, generalize to the level of individual students. Ecological correlations based on aggregate data are generally inflated estimates of lower level relationships (Robinson, 1950; Thorndike, 1939). Similarly, when micro researchers attempt to generalize findings from individual-level studies to higher levels, they commit an atomistic fallacy. Just because the relation holds at the lower level does not mean it will also hold at higher levels. Relationships that hold at one level of analysis may be stronger or weaker at a different level of analysis, or may even reverse direction (Ostroff, 1993).

Suppose, for example, that in his prior research Faust finds that individuals who report that their supervisors are high in charisma perform better than individuals who report that their supervisors are low in charisma. Although this individual-level relationship between perceptions of charisma and performance may hold within teams, it is entirely possible that the team-level relationship of charismatic leadership and performance is neutral and nonsignificant. Perhaps subordinates' perceptions of their leaders vary within teams but not between teams; team-level leader charisma may show little variability and hence fail to predict team performance. Or, the team-level relationship of charisma and performance may be negative. Perhaps highly charismatic team leaders are commonly assigned to low-performing teams in the hopes of enhancing team performance.

Should Faust then collect his individual-level data as planned and then simply aggregate the data to the team level, conducting team-level analyses of the data? Although some investigators do just this, we recommend against this practice because the meaning of measures aggregated in this way may be questionable. In the absence of careful theoretical work and subsequent statistical analyses designed to ensure the meaningfulness of the aggregated measures, Faust's team-level findings are likely to be highly misleading (James, 1982). Consider just two of Faust's variables. Faust's measures of charismatic leadership may, as we have suggested, vary considerably within teams, rendering the mean of team members' perceptions of their leader ambiguous at best. If so, their construct validity is questionable. If half of the members of a team describe their team leader as high in charisma whereas the remaining members of the team describe the team leader as low in charisma, the average of the team members'

ratings of the leader describes none of the team members' views. Furthermore, the average of team members' individual performance ratings may not constitute an effective measure of team performance; team members may perform well as individuals, but nevertheless function poorly as a united team. An orchestra nicely illustrates this possibility; the members of the orchestra may all be skilled musicians, but they may nevertheless sound terrible when they play together (Weick, 1998).

To explore the effects of team characteristics on individual performance, should Faust simply conduct individual-level analyses, after first collecting individual-level measures of the predictors (e.g., leader charisma), aggregating these measures to the team level, and then assigning the aggregated (average) scores back to individual team members such that the predictors are identical for the members of each team? As above, the answer is no. The issues of construct validity just discussed apply here as well. A team-level measure of charismatic leadership that lacks construct validity is no more meaningful as a predictor of individual performance than it is as a predictor of team performance. Furthermore, cross-level analyses in which, for example, team characteristics are used to predict individual-level outcomes raise complex theoretical and statistical issues resulting from the nonindependence of observations. (We discuss these issues in more detail below.) Cross-level analyses may be quite useful for Faust and for Acme, but an understanding of the theoretical and statistical issues arising in such analyses is essential for a valid interpretation of cross-level results.

In making a shift from micro to meso research, then, a researcher should step back to assess the relevance of his or her constructs to higher—more macro—levels of analysis. The researcher must then fashion a multilevel theoretical model depicting the hypothesized relationships among his or her constructs. Such theoretical work paves the way for data collection and analyses. These steps are examined in the following sections. Note that most of the variables of interest to Faust are meaningful, at least theoretically, at the team level of analysis. Thus, in shifting to encompass the team level, Faust need not alter his research focus substantially. In real life, our example is probably more the exception than the rule. We chose the variables of interest to Faust precisely because most were potentially relevant to the team-level of analysis.

Construct and Measurement Choices: What Is the Nature of Each Team Construct and How Should Each Construct Be Operationalized?

Faust's challenge is to provide Acme with information about team processes and team performance. Faust's individual-level constructs, measures, and results are very unlikely to illuminate team processes and team performance. Team-level analogues of the individual-level constructs of his original model may, however, do so. Certainly existing theory suggests that team leadership and team efficacy may affect team performance. Rigorous multilevel research rests not on such simple assertions, however, but on the careful definition, justification, and explication of the level of each focal construct in the model. In the absence of such precision and care, researchers may disagree about the nature, operationalization, and predicted effects of the construct (e.g., George, 1990; George & James, 1993; Yammarino & Markham, 1992). To help prevent the controversies and confusion that often surround the definition, meaning, and operationalization of team-level constructs, we distinguish three basic types of team- or higher level constructs: (a) global properties, (b) shared properties, and (c)

configural properties. Below, we define and discuss the measurement of the three types.

Global Team Properties

Global properties are relatively objective, descriptive, and easily observable team characteristics. Global properties, such as team function, characterize the team as a whole. Unlike shared and configural team properties, global team properties do not originate in or emerge from the characteristics of individual team members. Thus, global team properties have standing apart from member characteristics or social psychological processes. Consider team function, for example. Even if all the members of a company's customer service team turn over, the function of the team within the company is likely to remain the same. In contrast, team cohesion (a shared team property, described below) might well change dramatically. Of Faust's variables, only team location (that is, whether Acme's teams are face-to-face and proximal or virtual and geographically distributed) is clearly a global characteristic of the team.

Insofar as global properties are relatively objective, descriptive, and easily observable, there is typically no need to collect data from all of the members of a team to assess a global team property. Rather, a single expert individual is likely to be the appropriate source of data measuring a global construct. Acme's president, for example, might report the location of each team. Alternatively, each team leader might report his or her team's location. Global properties are typically far easier to measure than are shared and configural properties.

Shared Team Properties

Properties of this type originate in experiences, attitudes, perceptions, values, cognitions, or behaviors that are held in common by the members of a team. Examples of shared team properties include team cohesion, team norms, team climate, and team mental models. In postulating that a given variable is a shared team property, theorists and researchers ideally explain how and why team members come to share the construct of interest: What factors or processes constrain variability among team members, rendering the construct a shared property of the team? Common explanatory processes include homogeneous organizational context factors; attraction, selection, attrition; socialization; leadership; and social interaction (James & Jones, 1974; Klein, Dansereau, & Hall, 1994; Kozlowski & Doherty, 1989; Rentsch, 1990; Schneider & Reichers, 1983).

Of Faust's variables, both leadership perceptions and efficacy might constitute shared team properties. Are leadership perceptions shared among the members of a team? This question strikes at the heart of a long-running debate within leadership theory and research (e.g., Dansereau & Yammarino, 1998). Many leadership scholars suggest, implicitly or explicitly, that team members are likely to be consistent (homogeneous) in their perceptions of the leader. A leader, these scholars suggest, is likely to treat his or her subordinates in a similar and consistent fashion, causing subordinates to describe their leader in similar terms. Furthermore, interaction among team members—that is, social information processing—is likely to render team members homogeneous in their perceptions of the leader. Other leadership scholars counter that leaders are likely to differentiate among their subordinates, treating some subordinates as

in-group members and others as out-group members. This suggests that leadership perceptions are unlikely to be shared among the members of a team; some team members may perceive the team leader to be far more considerate or charismatic, for example, than may other team members. This example underscores the importance of precise theoretical specification and justification of the level and nature of each construct. Faust should theorize how and why team members' leadership perceptions are or are not shared within each team. His theoretical conclusions will inform his subsequent measurement, sampling, and analytical choices.

Efficacy, another of Faust's variables, might also be conceptualized as a shared team property. However, team efficacy (i.e., team members' belief in the efficacy of the team as a whole) seems far more likely than self-efficacy (individual's belief in their own individual efficacy) to be shared among team members. Chan (1998) refers to this distinction between these types of higher level constructs as "referent shift consensus" and "direct consensus," respectively. Although individual team members are likely to vary in their confidence in their own individual ability (self-efficacy), they may, as a result of team interactions and shared experiences, agree in their perceptions of the team's ability (team efficacy). This example illustrates the extent to which constructs may shift in meaning as a researcher shifts levels of analysis. Team or collective efficacy, as we and others have defined it (e.g., Bandura, 1997; Kozlowski, Gully, Nason, & Smith, 1999), is quite different than team members' average self-efficacy. Recall the example of the orchestra: Orchestra members may rate their individual musical self-efficacy high, but nevertheless rate their team efficacy—their confidence in the ability of the orchestra as a whole—quite low.

To measure shared team properties, researchers gather data from individual team members. This data collection procedure allows the researcher to assess the extent to which constructs are indeed shared as predicted. If individual-level data do reveal substantial within-group agreement or homogeneity, individual-level data are aggregated to the team level of analysis to represent the shared team construct. In the absence of substantial within-group agreement, the data gathered to represent the shared team construct lack critical construct-level validity (James, 1982; Klein et al., 1994; Kozlowski & Hattrup, 1992). Existing theory and recent research suggest that the wording of survey items may influence the extent of within-group agreement and between-group variability in a survey measure (Klein, Conn, Smith, & Sorra, in press), or the strength of the relationship between aggregated constructs (DeShon et al., 1999). More specifically, items with group rather than individual referents may increase the within-group homogeneity and between-group variability typically expected of measures of shared team constructs. There is, however, considerable research demonstrating high consensus for items with individual referents. Moreover, theory must be the driving force in dictating the wording of survey items.

This brief review suggests that if Faust conceptualizes team leader charisma and team efficacy as shared team constructs, he should measure individual team members' perceptions of their leader's charisma and individual team members' confidence in the team. After collecting these measures, Faust should assess the extent of within-team agreement on these measures. If justified, Faust should then aggregate these measures to the team level of analysis to represent the shared team constructs.[1] As is discussed in more detail in a later section, existing analytical procedures provide related but distinct ways and criteria to justify such aggregation.

Configural Team Properties

Like shared team properties, configural team properties originate in, or emerge from, individual team members' experiences, attitudes, perceptions, values, cognitions, or behaviors. Configural team properties, however, capture the array, pattern, or variability of individual characteristics within a team. Configural team properties have received relatively little attention within the organizational literature. Examples include team interpersonal network density (Brass & Burkhardt, 1993), team personality composition (Barry & Stewart, 1997), and team age diversity (Tsui, Egan, & O'Reilly, 1992). In studying these and other configural team properties, researchers make no assumption that the individual characteristics of interest —here, interpersonal relationships, personality, age—are held in common by the members of the team. Rather, the researchers strive to capture the array or configuration of these individual characteristics within the team.

Of Faust's constructs, perhaps both non–work responsibilities (e.g., childcare, elder care) and team performance might be conceptualized as configural team constructs. Consider, first, non–work responsibilities. Some teams may be characterized by relatively high and homogeneous non–work responsibilities; all team members may be young parents, carrying substantial non–work responsibilities. Other teams may be characterized by relatively low and homogeneous non–work responsibilities; most or all team members may be young and single and may thus experience few non–work responsibilities. Finally, some teams may be composed of individuals who vary substantially in the magnitude of their non–work responsibilities.

Team performance is a configural team property insofar as team performance emerges from the complex conglomeration of individual team members' performance. Depending on the nature of the team task, team performance may reflect the following: the sum of individual team members' contributions (the better each individual performs, the better team performance); the poorest team member's contribution (the team can perform no better than its weakest performer); the best team member's contribution (one team member can carry the team to a high level); the variability of team members' contributions (the more varied team members' contributions, the greater the number of nonredundant resources available to the team); or some more complex combination of team member contributions.

The challenge for Faust is to determine and explain why a given individual-level characteristic (such as non–work responsibilities or individual performance) should be conceptualized, at the team level of analysis, as a configural team property and how this property should be operationalized. Possible operationalizations, as we have hinted, include the sum of individual team member values, indices of variability among team member values, the minimum or maximum value among a team's members, and measures of the team network (e.g., density, homophyly). The appropriate operationalization of configural team construct depends, of course, on the guiding theoretical conceptualization (see Kozlowski & Klein, 2000, for a more extensive discussion of emergence and configural constructs).

To assess configural team constructs, researchers must typically gather data from (or about) individual team members. Thus, for example, Faust must assess each individual team member's non–work responsibilities to operationalize the configuration of non–work responsibilities within each team. In some cases, researchers may gather

data regarding each individual from a single individual or archival source to operationalize configural constructs. For example, a researcher might obtain data regarding each individual team member's pay from public sources or from the director of human resources to operationalize team pay dispersion, a configural construction (e.g., Bloom, 1999). In sum, individual-level data are critical for measuring both configural and shared team properties, but only shared team properties require the demonstration of within-group consensus or consistency.

Model Choices: What Kind of Model Describes the Predicted Relationships Among the Constructs?

Multilevel constructs within a given realm, or nomological network, may be combined in a variety of different ways to create models differing in structure and focus. Here we outline three broad classes of models: single level, cross-level, and homologous multilevel models. In a subsequent section, we show how the type of model influences data sampling and the selection of an appropriate data analytic system

Single-Level Models

Single-level models describe the relationship among variables at one level of theory and analysis. Micro-organizational researchers are likely to find single, individual-level models most familiar. But, team-level models are also single-level models. Team-level models—models specifying the relationship among global, shared, and/or configural team constructs only—are far more complex, from a levels perspective, than are individual-level models. Although single team-level models do not include individual-level or organizational-level constructs, single team-level models may present a number of theoretical and data analytic challenges. Suppose, for example, that Faust predicts that team location (a global construct) moderates the influence of team charismatic leadership (a shared construct) on team performance (a configural construct). The resulting single-level model requires a strong theoretical rationale and justification to support the conceptualization and operationalization of team charismatic leadership and team performance. However, once the emergent constructs are raised to the team level, the team-level model is straightforward to test using, for example, hierarchical regression.

Cross-Level Models

Cross-level models describe the relationship among variables at different levels of analysis. Three types of cross-level models are most common. A cross-level, direct effects model suggests that a predictor variable at one level of analysis influences an outcome variable at a different—typically lower—level of analysis. Thus, for example, Faust might predict that team location (a group-level, global construct) has a direct effect on team member performance (an individual-level outcome variable). Or, Faust might predict that team leader charismatic leadership (a group-level, shared construct) has a direct effect on team member self-efficacy (an individual-level outcome variable). Note that cross-level models of this type can only explain between-unit (here, between-team) variability in individual-level outcome measures. Team location, varying between teams but not within teams, cannot explain within-team variability in indi-

vidual performance. Similarly, team charismatic leadership cannot explain within-team variability in self-efficacy.

Cross-level moderator models suggest that variables at two different levels of analysis (e.g., one group-level variable and one individual-level variable) interact to predict an outcome at the lower level of analysis (e.g., an individual-level outcome measure). Thus, for example, an individual-level variable might moderate the effects of a team-level variable on an individual-level outcome. Alternatively, a team-level variable may moderate the effects of an individual-level variable on an individual-level outcome. Mathematically, these two cross-level moderator models are equivalent. Faust might propose, for example, that team location moderates the relationship of individual team member self-efficacy and performance. Perhaps self-efficacy is a particularly strong predictor of individual performance when team members work at a distance from one another. Or, to express the same relationship in a different (although mathematically equivalent) way, perhaps the effects of team location on individual team member performance vary as a function of team-member self-efficacy. That is, working at a distance from one's team members may have a particularly detrimental effect on the performance of individual team members who are low in self-esteem.

The third type of cross-level model—the *cross-level frog-pond model*—describes the effects of individual group member's standing within a group on individual-level outcomes. The term *frog-pond* captures the comparison or ranking effect that is central to models of this type (Klein et al., 1994). Frog-pond models suggest, in essence, that the true or absolute size of a frog is irrelevant. What matters is a frog's relative size: Is it a big frog or a small frog compared to other frogs in the pond? Adopting a cross-level, frog-pond model, Faust might propose that an individual's performance relative to his or her team members' performance predicts individual self-efficacy. Thus, for example, an individual whose absolute, true level of performance is mediocre may have high self-esteem if he or she works with others who perform relatively poorly. Conversely, an individual whose performance is mediocre may have low self-esteem if the others in his or her team are high performers. In this example, we have made relative performance the predictor rather than outcome variable only because this seems to make our example more intuitively obvious. Although we have categorized frog-pond models as cross-level models because they incorporate both unit- and individual-level components, others (e.g., Dansereau & Yammarino, 2000; Klein et al., 1994) consider frog-pond models a distinct subset of multilevel models.

Homologous Multilevel Models

Models of this type specify that a relationship between two variables holds at multiple levels of analysis. Consider the relationship between dependence and power. At multiple levels of analysis (individuals, groups, organizations), an entity upon which a second entity is dependent has power over that entity. Faust might propose that the relationship between efficacy and performance is not a cross-level frog-pond relationship but a multilevel, homologous relationship. That is, at both the individual and team levels of analysis, efficacy is positively related to performance.

The primary value of homologous multilevel models is that they allow the researcher to generalize both constructs and functional relations linking the constructs across different levels of the organizational system. Thus, such models hold the promise to enhance the generality and applicability of theory, and to better integrate macro

and micro models of organizational behavior. One of the challenges of developing homologous multilevel models, however, is that in the effort to find constructs and functional relations that hold at multiple levels, one may so abstract and simplify the phenomenon of interest that it is no longer useful. Perhaps this is why there are so few good examples of homologous multilevel theoretical models in the literature (e.g., Lindsley, Brass, & Thomas, 1995; Staw, Sandelands, & Dutton, 1981), and virtually no good empirical tests of such models. At their best, multilevel homologous models are powerful, integrative and parsimonious. At their worst, however, models of this type are obvious and trite (Klein, Cannella, & Tosi, 1999).

Sampling Choices: What Kind of Sample Is Necessary to Operationalize Key Constructs and to Test Predicted Relationships?

Researchers experienced in testing single-level models know, of course, that their samples must contain sufficient variability to minimize the risk of range restriction. One does not, for example, test the effects of wealth in a sample of multimillionaires, nor the effects of height in a sample of professional basketball players. Sampling issues within multilevel research are more complex, but much the same. The researcher's challenge is to gather a sample containing sufficient between-unit variability to assess the effects of unit differences and, when testing the effects of shared unit properties, sufficient within-unit homogeneity to warrant aggregation of lower level data to the unit level.

Consider Faust's new research at Acme. Between-team variability in the measures of team-level properties is essential or Faust's findings will be inconclusive because of range restriction in the measures. As organizational context may constrain team inputs, processes, and outcomes, there is at least some possibility that measures of these constructs may be restricted in range. The fact that Acme's president wants Faust to study the correlates of team performance suggests that there is substantial variability in team performance and its correlates, but this is an assumption that Faust should examine in detail. For example, Faust might assess the extent of variability in archival team performance data or he might explore the extent to which company policies and practices (e.g., performance appraisal and reward systems) are likely to limit between team-variability in performance and processes.

Ideally, measures of shared team constructs show not only between-team variability, but also within-team homogeneity. In planning his research, then, Faust should be mindful of the existence and magnitude of potential sources of such homogeneity. Moreover, these sources must engender homogeneity within teams, but not across teams. Shared experiences and interactions within a team may foster such within-team homogeneity and between-team variability. In contrast, a strong organizational culture and consistently applied organization-wide policies may foster within-organization homogeneity—that is, homogeneity within and across teams.

Yet another possibility is that homogeneity is high in some teams but limited in others. Perhaps, for example, some of Acme's teams show substantial within-team homogeneity in team efficacy whereas other teams do not. Face-to-face interaction in real teams may engender such homogeneity, whereas e-mail and telephone interaction—in virtual teams—does not (Bell & Kozlowski, in press). Faust may test this proposition by examining the relationships between team location (a global construct) and

within-team homogeneity of team efficacy (a configural construct). In this scenario, homogeneity is not a prerequisite for aggregation, but a meaningful dependent variable in its own right (Brown, Kozlowski, & Hattrup, 1996; Klein et al., in press).

To operationalize configural constructs, a researcher seeks a sample in which the pattern or array of data varies from unit to unit. One can only test the correlates of within-unit variability (a configural construct), for example, if units vary in their within-unit variability. Earlier we noted that Faust might conceptualize team members' non–work responsibilities as a configural construct. Note, however, that Faust will not find a significant relationship between the configuration of team members' non–work responsibilities and team performance, as proposed earlier, unless the extent and array of team members' non–work responsibilities vary between teams.

Our discussion thus far has focused on the sampling implications of differing kinds of constructs. Also relevant is the nature of the researcher's model. Single-level models require no special sampling considerations beyond those we have already discussed. However, cross-level models do require special considerations, as they typically rest on an assumption of within-unit homogeneity and between-unit variability on some unit-level constructs and an assumption of within and between-unit variability (that is, between-individual variability within and across units) on other, individual-level constructs. Suppose, for example, that Faust predicts that team leader charisma (a shared team construct) moderates the relationship of self-efficacy and performance (two individual-level constructs). The ideal sample in which to test this cross-level moderator model is one in which (a) team leader charisma varies between- but not within-teams, and (b) both self-efficacy and performance vary both within- and between-teams. Cross-level frog-pond models also require both within-unit and between-unit variability on the predictor, frog-pond variable.

The sampling implications of multilevel homologous models are complex, in part because such models are more common in theory than in empirical research. Indeed, one can find few—if any—empirical examples of homologous models. One strategy to test such a model is to collect a different sample of data to represent each level of the homologous model. Thus, for example, a researcher might collect survey data from a sample of individuals to test the relationship between individual power and individual dependence. The same researcher might also collect survey data from a sample of organizations to test the relationship between organizational power and organizational dependence. Alternatively, a researcher may test a multilevel model within a single sample of data. For example, Faust might gather measures of individual efficacy, individual performance, team efficacy, and team performance at Acme and assess both the individual-level relationship of efficacy and performance, and the team-level relationship of team efficacy and team performance. In this example, Faust collects two different measures of efficacy (individual or team) and two different measures of performance, hoping that individual efficacy and individual performance both vary and co-vary within and between teams, whereas team efficacy and team performance both vary and co-vary between teams, but not within teams.

Analysis Choices: What Analytical Procedures May Be Used to Test the Predicted Relationships?

In this section, we offer a brief overview and comparison of the statistical procedures commonly used to analyze multilevel data. Our goal is not to identify the best

procedure to use, nor to advocate some procedures more than others. Rather, we hope to illuminate the conceptual underpinnings and distinctive features of each of the various procedures, clarifying the questions answered most effectively by each procedure. We first examine the procedures most commonly used to justify aggregating individual-level data to higher level units to represent shared group constructs. These procedures are r_{wg}, eta-squared, within and between analysis (WABA), and the two intraclass correlations, ICC(1) and ICC(2). We then examine the procedures used to test multilevel models. These procedures are WABA, cross-level operator analyses (CLOP), and hierarchical linear modeling (HLM). As WABA is used both to justify aggregation and to analyze the relationships between variables, we discuss WABA in the two subsections below. Our discussion of the various procedures is necessarily somewhat superficial. A detailed discussion of each procedure would far exceed the page limitations of this article. (For more information about these procedures see Bliese, 2000; Dansereau & Yammarino, 2000; and Hofmann, Griffin, & Gavin, 2000; and James & Williams, 2000).

Justifying the Aggregation of Lower Level Data to Higher Units of Analysis

Each of the procedures commonly used to justify aggregation provides an assessment of the extent to which lower level data (e.g., individual-level data) are homogeneous within units (e.g., groups), as assumed of shared unit-level constructs. As we explain below, each procedure provides a distinctive, yet often complementary, assessment of the extent of within-unit homogeneity of lower level data. Thus, Faust might use any or all of these procedures to assess the extent to which a shared team constructs—say, team efficacy—is indeed homogeneous, or shared, within teams.

The index r_{wg}. Developed by James, Demaree, and Wolf (1984), r_{wg} is unique among the procedures used to justify aggregation insofar as r_{wg} assesses the extent of consensus, agreement, or within-unit variability within a single unit for a single measure—a construct by group approach. All of the remaining procedures contrast within-unit and between-unit variability across an entire sample of units—a construct by sample approach. The r_{wg} index compares the variability of a given variable within a specific unit (here, team) to an expected variance (EV). If the variability within a unit is substantially smaller than the variability expected by chance, then the resulting r_{wg} value suggests that it is justifiable to aggregate lower level data, for the specific variable and specific unit in question, to the unit level of analysis. Values for r_{wg} are expected to range from 0 to 1, but may be negative (for the single-item estimator) or greater than 1 (for the multi-item estimator) if the variability within a unit exceeds the EV (Brown et al., 1996).

The EV or variability expected by chance is often operationalized as the variance of a rectangular distribution. This is a statistical definition of purely random responding, because each response category has an equivalent likelihood of being endorsed. Psychological responses, however, are subject to a variety of response biases and are therefore rarely random. This means that r_{wg} values based on an EV estimated from a rectangular distribution will typically overstate true agreement. James et al. (1984) explicitly noted this problem and suggested alternative ways to operationalize the EV when the assumption of purely random responding is not warranted—as it rarely is.

For example, Kozlowski and Hults (1987) used the response distribution from an independent sample responding to the same measure to pose an alternative estimate of the expected variance. In effect, they assumed that this response distribution from an independent sample represented nothing but response bias. This provided a lower bound estimate of within group agreement. Thus, true agreement was assumed to range between an upper bound estimate based on the rectangular EV and the lower bound estimate based on the alternative EV (Kozlowski & Hults, 1987). Another alternative to the use of an EV based on a rectangular distribution includes random group resampling (Bliese, 2000). This technique essentially ignores the group structure in the data and randomly creates pseudo groups. Multiple resampling passes allow the estimation of an EV based on the response distribution of the pseudo groups. The technique also allows for a test of the significance of the resulting r_{wg} value. Although this technique effectively addresses one of the most serious ambiguities of the r_{wg} index—estimating the EV—it also moves r_{wg} closer to the other indices that compare within group to total variance.

In using r_{wg} to justify the aggregation of individual-level measures of team efficacy to the team level, Faust would calculate the r_{wg} values for team efficacy for each team in his sample. Faust might then report the average or median r_{wg} value, and the range of r_{wg} values, for the teams in his sample. Common practice is to conclude that aggregation of individual-level measures to the team level is appropriate if the mean equals or exceeds .7. This threshold is merely a rule of thumb, however. Random group resampling provides a way to test the statistical significance of r_{wg} values (Bliese, 2000). If the r_{wg} value for some teams is less than .70, Faust might exclude those teams or measures from team-level analyses. Alternatively, Faust might rely on the mean r_{wg} value of .7 for the entire sample to justify aggregation of the team-efficacy data to the team level. However, such an approach undermines the value of assessing r_{wg} within each group to ensure that the construct is not misspecified in any team.

In sum, r_{wg} is designed to answer the following question: How high is within-group agreement on a given variable within a given unit? The r_{wg} index assesses within-unit variability uniquely. It does not assess within- versus between-unit variability in a given measure, as do the remaining procedures used to justify aggregation. Note that agreement within-groups may be quite high even if group means do not differ—that is, even if between-group variability is so low that the procedures described next would not support aggregation to the group level (James et al., 1984).

Eta-squared. The eta-squared value from a one-way analysis of variance provides an estimate of the extent to which individual-level variability on a given measure is explained by higher level units. Thus, for example, an eta-squared value of .25 suggests that a researcher can explain 25% of the total, individual-level variance in a given measure simply by knowing what unit, or team, each individual belongs to. Or, to put it another way: 25% of the variance in the measure is between-teams whereas the remaining 75% of the variance is within teams. Ultimately, the greater between-group variance relative to within-group variance, the larger the value for eta-squared. Eta-squared, as these interpretations suggest, rests on a comparison of the within-unit and between-unit variability in a single measure across an entire sample. Accordingly, eta-squared values may be statistically significant primarily because the unit means on the measure of interest vary greatly (between-unit variability is substantial), or primarily because units are quite homogeneous (within-unit variability is minimal), or both.

Furthermore, eta-squared may be statistically significant even if some teams show substantial within-team variability. Eta-squared values summarize the relative amount of between-team versus within-team variance across an entire sample of teams. Eta-squared, unlike r_{wg}, provides no team by team analysis. Indeed, all of all the consistency based approaches to assessing within group homogeneity (i.e., eta-squared, ICC (1), ICC (2), and WABA) provide no team-by-team analysis. Thus, the consistency based approaches have the potential to allow some degree of misspecification error. That is, they may allow measures to be aggregated to the unit level for some teams even when the construct is not shared within those teams. However, the extent to which this is a serious problem is unknown.

Faust might use eta-squared to assess the extent of within- versus between-team variability in team efficacy. Eta-squared answers the question, to what extent does a measure vary between-units versus within-units? The *F* test for the eta-squared reveals the statistical significance of the between-unit variability. Note that significance values are affected by sample size. Other things being equal, the larger the sample of individuals, the more likely eta-squared is to be statistically significant.

Furthermore, as in all regression and analysis of variance, the larger the number of predictors, the larger the resulting explained variance (here, eta-squared). Thus, other things being equal, eta-squared values resulting from a one-way analysis of variance of a given measure drawn from a sample of 200 individuals grouped into 10 teams of 20 people each will be smaller than eta-squared values for the same measure drawn from a sample of 200 individuals grouped into 40 teams of 5 people each. The first sample results in an equation with 9 predictors (10-1), whereas the second results in a sample with 39 predictors (40-1). Bliese and Halverson (1998) have criticized the use of eta-squared, demonstrating that eta-squared provides inflated estimates of between-unit variability especially when group sizes are small (less than 25 people per group).

Intraclass correlation (1). The intraclass correlation coefficient (1) is calculated using the results of a one-way analysis of variance (i.e., eta-squared) and is commonly interpreted in much the same way as eta-squared. Thus, ICC(1), like eta-squared, provides an estimate of the proportion of the total variance of a measure that is explained by unit membership (Bliese, 2000). ICC(1) may also be interpreted as an estimate of the extent to which raters are interchangeable—that is, the extent to which one rater from a group may represent all the raters within the group. The larger ICC(1), the more alike the raters are (James, 1982).

ICC(1) and eta-squared differ, however, insofar as ICC(1) values do not vary as a function of group size (Bliese & Halverson, 1998). Eta-squared and ICC(1) provide quite similar estimates of between-unit variability in samples composed of relatively large groups (i.e., groups of 25 individuals or more). In samples composed of smaller groups (i.e., groups of fewer than 25 individuals per group), however, ICC(1) provides much smaller estimates of between-unit variability than eta-squared (Bliese & Halverson, 1998). Like eta-squared, ICC(1) answers the question, to what extent does a measure vary between-units versus within-units? When group sizes are small, eta-squared and ICC(1) may, however, provide quite different estimates of between-group variance—that is, of the variance explained by group membership.

This discussion suggests that if Faust calculated both eta-squared and ICC(1) values for team efficacy in the Acme sample of 100 teams of 10 people each, the resulting eta-squared value would exceed the resulting ICC(1) value. Note, however, that the *F*

test for ICC(1) is identical to the *F* test for eta-squared. Thus, although Faust's eta-squared value for team efficacy would exceed his ICC(1) value for the same measure, the two values would have the same significance level. Researchers using eta-squared or ICC(1) to justify aggregation usually conclude that aggregation is justified when the *F* test for these values is significant. Faust—with a sample of 1,000 (assuming a perfect response rate)—would be very, very likely to obtain a significant *F*.

Intraclass correlation (2). The intraclass correlation (2), or ICC(2), answers the following question: How reliable are the group means within a sample? ICC(2) values, like other measures of reliability (e.g., Cronbach's alpha), are commonly considered acceptable if they equal or exceed .70. Mathematically, ICC(2) is a variation of ICC(1). Essentially ICC(2) is a function of ICC(1) adjusted for group size. Indeed, if one knows the ICC(1) value for a given measure within a given sample, and one knows the average size of the groups within that sample, one can use the Spearman-Brown formula to estimate ICC(2) (Bliese, 2000). Other things being equal, the larger the group size, the larger ICC(2). For example, if ICC(1) is .20 and the average group size in the sample is 5, then the expected value for ICC(2) is .56. Conversely, if ICC(1) is .20 and the average group size in the sample is 20, then the expected value for ICC(2) is .71.

The underlying logic here is that group means based on many people per group are more reliable—more stable—and thus more useful measures of group properties than are group means based on fewer people per group. If one imagines groups of extreme sizes, the point is intuitively obvious: The mean weight of, say, 500 adults is a highly stable or reliable number. Replace one half of the group's members and the mean weight is unlikely to change substantially. The mean weight of a group of just two individuals is utterly unstable. Replace one half the group (one of two people), and the mean weight of the two-person group may change substantially.

ICC(2) values will exceed .70 if the between-group variability of a measure is large, if group sizes in the sample are large, or both. In Faust's case, given groups of 10 people (again, assuming a perfect response rate), ICC(1) must equal or exceed .19 if ICC(2) is to equal or exceed .70.

WABA. Like ICC(1) and ICC(2), WABA builds on eta-squared, and resulting estimates of within-unit and between-unit variability. WABA differs from eta-squared in at least three critical respects, however. First, eta-squared, ICC(1), and ICC(2) are commonly used to differentiate between: (a) measures that show sufficient between-unit variability and within-unit agreement to justify aggregation to the unit level; and (b) measures that do not show sufficient between-unit variability and within-unit agreement to justify aggregation. WABA, in contrast, recognizes not two but three alternatives: (a) the measure shows sufficient within-unit agreement and between-unit variability to justify aggregating the measure and using it to operationalize a shared-unit construct (wholes); or (b) the measure shows sufficiently within-unit variability and sufficient between-unit agreement to justify using the measure to operationalize a frog-pond variable (parts); or (c) the measure varies both between-units and within-units and thus should be used to operationalize individual-level constructs (equivocal). Thus, whereas researchers using other procedures to justify aggregation might disregard evidence of substantial within-unit variability as between-unit variability or error, WABA users interpret such within-unit variability as evidence of a frog-pond or parts condition. Note, however, that other procedures also

provide evidence of such within-unit variability. For example, ICC (1) may assume negative values when within-unit variability is substantial. Similarly, r_{wg} values will exceed 1.00 (for the multiple-item estimator) or will be negative (for the single-item estimator) when within variance is greater than the expected variance (Brown et al., 1996). Such values are indicative of subgroup polarization or potential frog-pond phenomena (Brown & Kozlowski, 1999). WABA researchers simply devote greater attention to the parts condition than do most researchers using other statistical procedures to justify aggregation.

Second, WABA differs from other statistical procedures insofar as WABA provides, and rests on, tests of both statistical and practical significance. Thus, a researcher using WABA standards must show that a measure's between-unit variability is both statistically and practically significant prior to aggregating the measure to the unit level. WABA researchers use the *F* test to assess statistical significance—the same *F* test used by researchers examining eta-squared and ICC(1). WABA researchers use the *E* test (see Cohen, 1988; Dansereau, Alutto, & Yammarino, 1984) to assess practical significance. The *E* test provides a geometric assessment of the magnitude of between-unit variability versus within-unit variability. To meet WABA's strictest standards of practically significant between-variance (indicative of a "wholes" variable), between-unit variance must equal or exceed .75. (WABA also specifies two alternative—and more liberal—standards for practical significance. The most liberal standard suggests that the eta-squared for a wholes variable must exceed .50.) The *E* test is perhaps the most controversial element of WABA. Bliese and Halverson (1998) suggest that because the *E* test relies on eta-squared, the *E* test may be biased by group size, providing inflated estimates of the practical significance of between-group variability when group sizes are small.

Finally, WABA is designed to provide not just a test of the appropriateness of aggregation, but rather an overall assessment of the level of analysis most appropriate for analyzing the relationships among a set of variables within a given sample. Thus, researchers use WABA to assess the extent of within-unit and/or between-unit variance and covariance among all the variables in a sample. The first part of WABA (WABA I) examines each variable in a researcher's data set in isolation. The second part of WABA (WABA II, discussed below) examines the relationships among the variables in a sample.

In sum, Faust might use WABA to assess the statistical and practical significance of between-group variability in his measures. His findings of statistical significance—based on the *F* test—would equal the results he would obtain using eta-squared or ICC(1). A finding of statistical significance, however, does not guarantee a finding of practical significance. Thus, Faust might find between-unit variability in team efficacy to be both practically and statistically significant, neither practically and statistically significant, or either practically or statistically significant but not both. Faust would use WABA I to assess within- and between-group variability in all of the measures of his model, not just team efficacy. Thus, WABA I answers the question, "Do the measures in a data set show statistically and practically significant between- unit or within-unit variability?", as a step on the way to answering another question, "At what level or levels should the variables in a data set by analyzed?"

A brief summary. In unambiguous cases—when a measure's variance lies quite predominantly between-units, or quite predominantly within-units, or roughly equally

within- and between-units—the five statistical procedures we have outlined are likely to lead to the same conclusion. In more ambiguous cases, however, the statistical procedures may lead to differing conclusions regarding the appropriateness of aggregating data to the unit level. In such cases, a researcher should rely on theory and prior research to guide the selection of an appropriate statistical justification to inform his or her aggregation decisions.

Analyzing the Relationships Among Variables: Testing Single-Level Relationships, Multilevel Homologous Relationships, and Cross-Level Relationships

WABA, CLOP, and HLM are all designed to test the relationships among multilevel data. As above, however, the procedures differ in approach and focus; they answer different questions and are thus most applicable to different types of multilevel models. Below, we outline these differences. As above, given space limitations, our presentation cannot do justice to the details and nuances of each procedure.

WABA. As suggested above, WABA I, is designed to assess the extent of within-unit and between-unit variance in a single measure. WABA II is designed to assess the extent to which two or more variables covary primarily within-units, between-units, or both within- and between-units. Together, WABA I and WABA II are designed, as we have noted, to answer the following question: At what level or levels should the variables in a data set be analyzed? WABA is most effective in testing single-level models and multilevel homologous models.

The correlation between any two measures drawn from a sample of individuals across units may reflect between-unit covariance in the measures, within-unit covariance in the measures, or both between-unit and within-unit covariance in the measures (Dansereau & Yammarino, 2000). WABA II rests on this fundamental insight. Suppose, for example, that the raw score, individual-level correlation of team leader charisma and team efficacy is .40. This correlation may reflect between-unit variance and covariance in the measures: Mean scores for each measure vary, and covary, from team to team; but within teams, the relationship of the two measures is weak. If so, the two measures share, in WABA parlance, a "wholes" relationship—a team-level relationship of two shared team constructs. Accordingly, WABA researchers would aggregate the measures to the team level and calculate their team-level relationship.

Alternatively, the .40 correlation of charisma and efficacy may reflect within-unit variance and covariance in the measures: Scores for each measure vary and covary substantially within each team, but the team-level correlation—the correlation between the team means for the two measures—is weak. If so, the two measures form, in WABA parlance, a "parts" relationship: a within-team relationship of two frog-pond constructs. WABA researchers would calculate the within-group deviation scores for each measure and report the within-group relationship of the measures.

Finally, the .40 correlation may reflect both within-unit and between-unit variance and covariance: Scores for each measure vary and covary within-teams and further mean scores for each measure vary and covary between teams. If so, the two measures have an "equivocal" relationship, in WABA parlance. The relationship is not a wholes relationship, distinguished by statistically and practically significant between-team

variance and covariance of the measures. Nor is the relationship a frog-pond relationship, distinguished by statistically and practically significant within-team variance and covariance of the measures. Rather, the variance and covariance of the two measures within and between teams suggests that the relationship of the measures is best conceptualized as an individual-level relationship reflecting individual differences in team leader charisma and team efficacy. Accordingly, WABA researchers would simply analyze the raw score individual-level relationship of the two measures.

To discern which of these three possibilities best describes the relationship of two variables, WABA researchers use the *A* test and *Z* test to analyze the within- and between-unit components of the raw score correlation (Dansereau et al., 1984). Typically WABA researchers assess the within-unit and between-unit variance and covariance of all the measures in a data set. WABA researchers then test the relationships between the independent and dependent variables at the single level of analysis that best characterizes the data set as a whole.

Thus, if Faust used WABA I and II, he would analyze all of his data, examining the extent to which all of the measures vary and covary between- and within-teams. WABA is not designed to test cross-level direct effect models, thus Faust could not test the simultaneous effects of, say, team efficacy (wholes) and self-efficacy (equivocal) on performance (equivocal). Rather, Faust would use WABA to test single-level relationships in the data. If he found that most or all of the measures varied and covaried both within- and between-teams, he would—following WABA standards—interpret the relationship as an individual-level relationship. If, instead, he found that most or all of the measures varied and covaried predominantly between-teams, he would interpret the relationship as a team-level relationship. (Given the focus of this research, Faust would be unlikely, it seems, to find predominant variance and covariance within teams.) Faust might conceivably split his data set, analyzing and interpreting some relationships at the individual level of analysis, and others at the team level of analysis.

WABA, we have emphasized, is primarily designed to test at what level of analysis the relationships within a data set should be tested. Homologous, multilevel models suggest that the relationship between two or more variables is isomorphic (that is, parallel or identical) at two different levels of analysis. WABA may be used to test a homologous, multilevel model. For example, the relationship between efficacy and performance may be homologous—occurring at the individual level of analysis, the team level of analysis, and the organizational level of analysis. WABA might be said to support such a conclusion if individual-level measures of efficacy and performance were found to vary and covary between teams, forming a team-level wholes relationship, and team-level scores (means) for efficacy and performance were found to vary and covary between organizations, forming an organization-level wholes relationship. (Note that Faust would need data from other organizations too, not just Acme, to test such a model.) However, if individuals are homogeneous within teams, then the individual-level relationship may not reflect individual differences, but team differences. Similarly, if teams are homogeneous within organizations, then the team-level relationship may not reflect team differences, but organizational differences.

To document a homologous, multilevel relationship between efficacy and performance, one might, as we noted earlier, gather two measures of efficacy (self- and team-efficacy) and two measures of performance (individual performance and team performance). One could then use WABA to explore the levels of the efficacy-performance relationship. Ideally—at least from a WABA perspective—the results would show that

self-efficacy and individual performance have an equivocal relationship, whereas team efficacy and team performance have a wholes relationship.

CLOP. CLOP encompass a number of analysis of variance, covariance, and regression approaches designed to examine the main or direct effects of higher level (e.g., team) variables on lower level (e.g., individual) outcomes and/or the moderating effects of higher level variables on lower level relationships (James & Williams, 2000). Accordingly, CLOP answers the following questions: What is the effect of a higher level unit characteristic on a lower level outcome? and/or, What is the influence of a higher level unit characteristics on the relationship between lower level variables?

The logic of CLOP is quite straightforward. If, for example, Faust wanted to assess the effects of team location (a group-level, global construct) and team leader charisma (a group-level, shared construct) on team member performance (an individual-level outcome variable), Faust would disaggregate the team location dummy code and mean team scores for team leader charisma to the individual level of analysis. Thus, each individual in a given team would have the same team location scores (indicating that his or her team is virtual or that it is real) and the same team leader charisma score. Faust would then use regression (i.e., general linear modeling) to determine the amount of variance in team member performance attributable to team location and to team leader charisma.

Three details regarding this approach bear mention. First, before using CLOP, researchers must use r_{wg}, eta-squared, ICC(1), ICC(2), and/or WABA to justify the aggregation of data to the unit level (to operationalize unit-level constructs). Unlike WABA, CLOP itself incorporates no tests designed to ascertain the extent to which measures are shared or homogeneous within units. Second, by definition, predictor variables varying solely between units (i.e., unit-level predictor variables) can only explain between-unit variance in outcome measures. Thus, if team leader charisma (a shared team-level variable, varying solely between and not within teams) explains 15% of the variance in individual performance, it actually explains a larger percentage of the between-team variance in performance while explaining none of the within-team variance in performance. Finally, CLOP may be used to examine the combined effects of team-level and individual-level predictors on individual-level outcomes. For example, Faust might examine the combined effects of team leader charisma (again, a shared team-level variable disaggregated to individual team members) and individual self-efficacy on individual performance. To do so, Faust would simply include the two independent variables in his regression equation.

CLOP may also be used to examine the interactive effects of higher (e.g., team) and lower (e.g., individual) level variables on lower level (e.g., individual) outcomes. For example, Faust might use CLOP to test the hypothesis that team location moderates the relationship of self-efficacy and performance. Perhaps the relationship between individual self-efficacy and performance is stronger in virtual teams than in real teams. To test this hypothesis, Faust would simply conduct a hierarchical regression first examining the direct effects of team location (disaggregated to team members) and self-efficacy on individual performance, and then examining the interaction effects of the two predictor variables on individual performance.

CLOP thus provides a relatively straightforward and flexible approach to test cross-level models. CLOP has been criticized, however, for two interrelated reasons. First, in testing the direct effects of a higher level (team) variable on a lower level (indi-

vidual) variable, CLOP estimates the standard error for the higher level variable based on total (e.g., within and between team) variance, rather than based on higher level (e.g., between-team) variance alone (Klein et al., 2000). The resulting standard error estimate for the higher level variable is smaller than it should be (Bryk & Raudenbush, 1992; Kreft & de Leeuw, 1998). As a result, tests of the statistical significance of the higher level variable in explaining variance in the outcome measure may be biased—that is, too liberal. Second, in testing the direct effects of a higher level variable on a lower level variable, CLOP attributes N-1 degrees of freedom to the higher level variable, where N equals the total number of observations—here, the total number of individuals. The appropriate number of degrees of freedom for the higher level variable, however, is J-1, where J equals the number of units or teams. As above, this may result in too-liberal tests of the statistical significance of the higher level variable in explaining variance in the lower level outcome variable. As we explain in more detail below, HLM and other random coefficient modeling techniques were designed, in part, to address these issues.

HLM. HLM is one of a class of several multilevel random coefficient modeling techniques. Several statistical software packages may be used to conduct multilevel random coefficient modeling. Because HLM is perhaps most familiar, we focus on it in this discussion. However, the issues discussed here are applicable regardless of which software package is used. Like CLOP, HLM is designed to test cross-level direct effect and moderating effects models. Thus, HLM—like CLOP—answers these important questions: What is the effect of a higher level unit characteristic on a lower level outcome? And/or, What is the influence of a higher level unit characteristic on the relationship between lower level variables? Furthermore, like CLOP, HLM includes no procedures to justify aggregation. Thus, researchers using HLM should use other procedures (i.e., r_{wg}, WABA, ICC(1), etc.) to justify the aggregation of individual-level data to the unit level, prior to testing the effects of such unit-level measures on lower level variables (e.g., Hofmann & Stetzner, 1996). As we explain below, Faust might use either CLOP or HLM to test the models we described in the preceding section. Were Faust to use both techniques to test the same models in the same data set, his results would be very similar, but not identical.

HLM is conducted as a simultaneous, two-stage process (Hofmann et al., 2000). In the Level 1 or first stage, HLM analyzes the relationship among lower level (e.g., individual-level) variables within each higher level unit (e.g., team), calculating the intercept and slope(s) for the lower level model within each unit. In the second step, HLM analyzes the relationship between higher level (e.g., team-level) variables and the intercepts and slopes for each team. In other words, Level 2 analyses treat variance in within-team slopes as indicative of moderation and variance in within-team intercepts as indicative of direct effects.

Suppose, as above, that Faust seeks to test a cross-level, direct effects model in which team location and team leader charisma are hypothesized to influence individual performance. In the first stage, HLM would calculate the intercepts for each team—that is, each team's mean individual performance score. In the second stage, HLM would calculate the relationship between the two higher level predictors—team location and team leader charisma—and the team performance intercepts. The results would document the extent to which team location and team leader charisma explain

between-team variance in individual performance (i.e., variance among the team mean scores for individual performance.)

Consider a somewhat more complex example. Suppose Faust wants to test the same model as described above, after controlling for the effects of individual self-efficacy on individual performance. In the first stage, HLM would calculate, for each team, the intercept and slope of the within-team relationship of individual self-efficacy and individual performance. In the second stage, HLM would examine the relationship between the team performance intercepts and the two higher level predictors (again, team location and team leader charisma). The results would describe the extent to which individual self-efficacy explains within-team variance in individual performance and the extent to which team location and team leader charisma explain between-team variance in individual performance.

Finally, suppose that Faust wants to test the model above, as well as the moderating effects of team location and team leader charisma on the relationship between self-efficacy and individual performance. HLM would examine the relationships specified above, as well as the relationship between the within-team slopes (specifying the relationship between self-efficacy and performance in each team) and the two higher level predictors. This final analysis would thus extend the preceding analyses, documenting the extent to which team location and team leader charisma explain variability in the slope of the self-efficacy—individual performance relationship. Perhaps, as we speculated earlier, this individual-level relationship is relatively strong in virtual teams and relatively weak in teams guided by a charismatic team leader.

CLOP and HLM may be used, as we have suggested, to test the same models. A researcher using both techniques to examine the same variables in the same data set would find the HLM parameter estimates and CLOP parameter estimates to be identical, or virtually identical. The statistical significance of the parameter estimates might differ, however. Furthermore, variance explained would differ. In testing the effects of a cross-level direct effect variable, CLOP reports the extent to which the higher level (here, team) variable explains total (here, between-individual) variance in the lower level variable. Thus, for example, CLOP might show that team leader charisma explains 15% of the total variance in individual performance. However, if only 20% of the variance in individual performance is between-team variance (and 80% is within-team variance), then HLM would report that team leader charisma explains roughly 75% (or 15/20) of the between-team variance in individual performance. The HLM results may appear, at first blush, to be far more powerful and impressive than the CLOP results. Who, after all, would not prefer to explain 75% of the variance rather than 15% of the variance? But, the results are essentially the same; only the metric is different. Accordingly, it is not difficult to translate CLOP results to approximate the results one would find in using HLM (see, James & Williams, 2000; Klein et al., 2000). Nevertheless, HLM is currently regarded as the more rigorous and advanced technique, allowing researchers to use appropriate standard errors and degrees of freedom and to thereby avoid potentially important violations of statistical assumptions of independence.

A brief summary. Clearly, CLOP and HLM are similar analytical strategies. Both are well designed to test cross-level direct effect and moderating models. WABA is quite different. It was designed to verify the appropriate level of analysis for a given set of variables in a given data set and to test single-level and multilevel homologous mod-

els. Thus WABA differs from CLOP and HLM in two essential respects. First, WABA includes procedures and standards to evaluate the level of variables and of relationships (including procedures and standards to justify the aggregation of lower level variables to higher levels of analysis): CLOP and HLM do not include such procedures. And second, WABA is typically used to assess single-level models, although it can also be used to test multilevel homologous models: CLOP and HLM, in contrast, are typically used to assess cross-level relationships, although clearly they can also assess single-level models.

Thus, researchers using CLOP, HLM, and WABA to analyze the same relationships in the same data set may reach quite different conclusions. First, WABA, r_{wg}, eta-squared ICC(1), and ICC(2) may lead to different conclusions about whether particular variables can be aggregated. In our experience, we have found that WABA imposes more stringent standards than the other statistical procedures often used to justify aggregation. Thus, WABA researchers might well advise against aggregating a measure that HLM and CLOP researchers, using other procedures to justify aggregation, might readily aggregate. Second, researchers using CLOP or HLM use variables at different levels of analysis (e.g., team- and individual-level predictors) to predict lower level variables (e.g., individual outcomes). WABA, in contrast, largely constrains researchers to assess single-level relationships. Although WABA can be used to test cross-level moderator relationships, it was not designed to do so and does so less flexibly than HLM and other multilevel random coefficient modeling techniques.

Klein et al. (2000) present an example illustrating these differences among the analytic techniques. The research team used different procedures—WABA, CLOP, and HLM—to analyze a simulated multilevel data set. Researchers using WABA chose not to aggregate any of the variables. Researchers relying on ICC(1) and ICC(2) values to justify aggregation chose, instead, to aggregate several variables. Furthermore, the WABA researchers conducted single-level (individual-level) analyses to assess the relationship between the predictors and job satisfaction, the simulated dependent variable. The CLOP and HLM researchers used cross-level modeling to test these relationships. The results highlight important differences between WABA, on one hand, and CLOP and HLM, on the other. The results also document differences between the results obtained using CLOP and those obtained using HLM. Thus, for example, a comparison of the HLM and CLOP results indicates that team-level predictors explain more variance in the HLM results than in the CLOP results. The key, of course, is that HLM reports between-team variance explained by team-level predictors whereas CLOP reports total variance explained by the team-level predictors. The CLOP and HLM results also differ slightly in the reported significance of the predictors. Bottom line, the CLOP and HLM researchers came to the same substantive conclusions; the findings of WABA researchers were distinct.

Conclusion

There is increasing interest in, and proliferation of, research that attempts to bridge the micro-macro gap by modeling phenomena that cut across multiple levels of analysis. For too long, micro-researchers have routinely neglected the effects of the organizational contexts within which individual behavior occurs. Organizations are hierarchically nested systems. To neglect these systems' structure in our conceptualization and research designs is to develop incomplete and misspecified models. Although multi-

level theory and multilevel research hold great promise for enhancing our understanding of organizational behavior, they also hold the potential for creating many pitfalls for the unwary researcher. The many vigorous debates in the literature regarding appropriate ways to conceptualize multilevel models, support decisions to aggregate data, and to evaluate multilevel models have done much to create confusion around the conduct of multilevel research (Kozlowski & Klein, 2000).

This article represents one small step in our efforts to dispel this confusion. We have traced the steps a researcher must take in planning and carrying out multilevel research. Multilevel research, we have emphasized, should begin with well-developed theory, and careful identification and definition of key constructs. Construct definition includes the specification and theoretical justification of the level and nature of the construct.

Multilevel research continues with the choice of the model specifying the relationship among the constructs. Hypotheses in multilevel research are level-specific. Thus, hypotheses describe not simply the direction—positive or negative—of the relationship between constructs but also the level or levels of each predicted relationship: single, cross-level direct, cross-level moderating, or multilevel homologous.

Construct and model choices drive the operationalization of the constructs—the measurement strategies—and the sampling plan. Researchers, we have noted, can and should make an effort to ensure that their measures show within- and between-variability across units as predicted by the theory.

When the researcher postulates the existence of shared constructs, great care is warranted in the justification for aggregation. Several statistical tools appropriate for supporting aggregation decisions were discussed and compared. There is no one universally preferred approach for assessing shared constructs.

Finally, a number of multilevel analytical systems are available to researchers. These systems are best suited to answering different research questions: There is no one single best way. Informed by an understanding of the logic and assumptions embedded within each analytic system, and the demands of their theoretical model, researchers can make appropriate choices among statistical approaches that best satisfy their requirements.

Multilevel research is challenging, but the keys to good multilevel research are simple: Detailed and well-developed theory; judicious operationalization, measurement, and sampling; appropriate justification for aggregation; and thoughtful selection of an analysis strategy. In other words, there is no substitute for the basics!

Note

1. Throughout this discussion, we have emphasized that the definition of shared team constructs presupposes within-team homogeneity. Thus, prior to operationalizing a shared team construct by averaging team members' individual responses, researchers should verify that team members indeed show within-team agreement. We should note, however, that in some cases, researchers may aggregate individual scores to the team (or other unit) level even though individuals' scores show little within-team (or within-unit) homogeneity. For example, a researcher might calculate team members' average age, or—outside of organizational research—the average birth weight for children in a city, state, or nation. In such cases, the variable measured is not a shared property of the unit, as we have defined the term; it is simply the unit average. Analyses of unit averages are acceptable statistically, but may raise problems of interpretation as it may be

unclear whether the unit average reflects the experiences of most or all of the members of a team, or instead reflects the influence of a few outliers.

References

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Barry, B., & Stewart, G. L. (1997). Composition, process, and performance in self-managed groups: The role of personality. *Journal of Applied Psychology*, *82*, 62-78.

Bell, B. S., & Kozlowski, S.W.J. (in press). Virtual teams: Implications for leadership. *Groups and Organization Management*.

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 consensus and psychological well-being: A large field study. *Journal of Applied Social Psychology*, *28*, 563-580.

Bloom, M. (1999). The performance effects of pay dispersion on individuals and organizations. *Academy of Management Journal*, *42*, 25-40.

Brass, D. J., & Burkhardt, M. E. (1993). Potential power and power use: An investigation of structure and behavior. *Academy of Management Journal*, *36*, 441-470.

Brown, K. G., & Kozlowski, S.W.J. (1999, April). Toward an expanded conceptualization of emergent organizational phenomena: Dispersion theory. In F. P. Morgeson & D. A. Hofmann, (Chairs), *New perspectives on higher-level phenomena in industrial/organizational psychology*. Symposium conducted at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

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. In S. Kozlowski & K. Klein (Chairs), *The meaning and measurement of within-group agreement in multi-level research*. Symposium conducted at the 56th Annual Convention of the Academy of Management Association, Cincinnati, OH.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, 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. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

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

Dansereau, F., & Yammarino, F. J. (1998). Introduction and overview. In F. Dansereau & F. J. Yammarino (Eds.), *Leadership: The Multiple-Level Approaches* (pp. xxv-xliii). Greenwich, CT: JAI Press.

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.

DeShon, R. P., Milner, K. R., Kozlowski, S.W.J., Toney, R. J., Schmidt, A., Wiechmann, D., & Davis, C. (1999, April). The effects of team goal orientation on individual and team performance. In D. Steele-Johnson (Chair), *New directions in goal orientation research: Extending the construct, the nomological net, and analytic methods*. Symposium conducted at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Firebaugh, G. (1979). Assessing group effects: A comparison of two methods. *Sociological Methods & Research*, *7*, 384-395.

George, J. M. (1990). Personality, affect, and behavior in groups. *Journal of Applied Psychology*, *75*, 107-116.

George, J. M., & James, L. R. (1993). Personality, affect, and behavior in groups revisited: Comment on appregation, levels of analysis, and a recent application of within and between analysis. *Journal of Applied Psychology*, *78*, 798-804.

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.

Hofmann, D. A., & Stetzer, A. (1996). A cross level investigation of factors influencing unsafe behavior and accidents. *Personnel Psychology*, *49*, 307-339.

House, R., Rousseau, D. M., & Thomas-Hunt, M. (1995). The meso paradigm: A framework for integration of micro and macro organizational. In L. L. Cummings & B. Staw (Eds.), *Research in organizational behavior* (Vol. 17, pp. 71-114). Greenwich, CT: JAI.

James, L. J., & Williams, L. (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.

James, L. R. (l982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology*, *67*, 2l9-229.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within group interrater reliability with and without response bias. *Journal of Applied Psychology*, *69*, 85-98.

James, L. R., & Jones, A. P. (l974). Organizational climate: A review of theory and research. *Psychological Bulletin*, *81*, l096-lll2.

Klein, K. J., Bliese, P. D., Kozlowski, S.W.J., Dansereau, F., Gavin, M. B., Griffin, M. A., Hofmann, D. A., James, L. R., Yammarino, F. J., & Bligh, M. C. (2000). Multilevel analytical techniques: Commonalities, differences, and continuing questions. In K. J. Klein & S.W.J. Kozlowski (Eds.), *Multilevel theory, research and methods in organizations: Foundations, extensions, and new directions* (pp. 512-553). San Francisco, CA: Jossey-Bass.

Klein, K. J., Cannella, A., & Tosi, H. (1999). Multilevel theory: Challenges and contributions. *Academy of Management Review*, *24*, 243-248.

Klein, K. J., Conn, A. B., Smith, D. B., & Sorra, J. S. (in press). Is everyone in agreement? An exploration of within-group agreement in employee perceptions of the work environment. *Journal of Applied Psychology.*

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., & Doherty, M. L. (1989). Integration of climate and leadership: Examination of a neglected issue. *Journal of Applied Psychology*, *74*, 546-553.

Kozlowski, S.W.J., Gully, S. M., Nason, E. R., & Smith, E. M. (1999). Developing adaptive teams: A theory of compilation and performance across levels and time. In D. R. Ilgen & E. D. Pulakos (Eds.), *The changing nature of work performance: Implications for staffing, personnel actions, and development* (pp. 240-292). San Francisco: Jossey-Bass.

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., & Hults, B. M. (l987). An exploration of climates for technical updating and performance. *Personnel Psychology*, *40*, 539-563.

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: Foundations, extensions, and new directions* (pp. 3-90). San Francisco, CA: Jossey-Bass.

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

Lindsley, D. H., Brass, D. J., & Thomas, J. B. (1995). Efficacy-performance spirals: A multilevel perspective. *Academy of Management Review*, *20*, 645-678.

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

Rentsch, J. R. (1990). Climate and culture: Interaction and qualitative differences in organizational meanings. *Journal of Applied Psychology*, *75*, 668-681.

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

Schneider, B., & Reichers, A. E. (l983). On the etiology of climates. *Personnel Psychology*, *36*, l9-39.

Staw, B., Sandelands, L. E., & Dutton, J. E. (1981). Threat-rigidity effects in organizational behavior: A multilevel analysis. *Administrative Science Quarterly*, *26*, 501-524.

Thorndike, E. L. (1939). On the fallacy of imputing the correlations found for groups to the individuals or smaller groups composing them. *American Journal of Psychology*, *52*, 122-124.

Tsui, A. S., Egan, T. D., & O'Reilly, C.A. (1992). Being different: Relational demography and organizational attachment. *Administrative Science Quarterly*, *37*, 547-579.

Yammarino, F. J., & Markham, S. E. (1992). On the application of within and between analysis: Are absence and affect really group-based phenomena? *Journal of Applied Psychology*, *77*, 168-176.

Weick, K. E. (1998). Improvisation as a mindset for organizational analysis. *Organizational Science*, *9*, 543-555.

Katherine J. Klein is an associate professor of industrial and organizational psychology at the University of Maryland. Her research interests focus on organizational and technological change, innovation implementation, multilevel theory and research, leadership, and team processes. She is a fellow of the Society for Industrial and Organizational Psychology and the American Psychological Association. Her articles have appeared in the Journal of Applied Psychology, *the* Academy of Management Review, Leadership Quarterly, *and the* Journal of Vocational Behavior. *She currently serves on the editorial boards of the* Academy of Management Review, *the* Journal of Applied Psychology, *and the* SIOP Frontiers Series.

Steve W. J. Kozlowski is a professor of organizational psychology at Michigan State University. His major interests are in the application of multilevel theory to understanding innovation, change, and adaptability at the individual, team, and organizational levels. His current work, which focuses on team leadership, training and development, and adaptive performance, has been supported by the Naval Air Warfare Center Training Systems Division and the Air Force Office of Scientific Research. His work has been published in many journals, including the Academy of Management Journal, Human Performance, Human Resource Planning, *and the* Journal of Applied Psychology. *He serves or has served on the editorial boards of the* Journal of Applied Psychology, Organizational Behavior and Human Decision Processes, *and the* Academy of Management Journal. *He is a fellow of the American Psychological Association and the Society for Industrial and Organizational Psychology.*